Karma, Kai & Komulainen, Erkki

Käyttäytymistieteiden tilastomenetelmien jatkokurssi

Toinen laitos

(Versio 2.2, 1.1.2002)

Helsingin yliopisto

Kasvatustieteen laitos

Sisällys

ISBN 952-10-0290-5 (Word)
ISBN 952-10-0291-3 (pdf)

Käyttäjälle

Käsillä oleva esitys kattaa sellaiset keskeiset alueet, jotka yleensä kuuluvat käyttäytymistieteiden aineopintoihin. Perusseikat, kuten keski- ja hajontaluvut sekä erityisesti korrelaatio oletetaan tunnetuiksi. Ne on usein syytä kerrata. Tällainen oppiaines on esim. Komulainen & Karma (2001) "Tilastollisen kuvauksen perusteet käyttäytymistieteissä" 2. laitos, joka on saatavana Kasvatustieteen laitoksen sähköisistä oppimateriaaleista (kuten käsillä olevakin teksti).

Ensimmäisen laitoksen esipuheeseen ei juuri ole täydennettävää. Harjoitusesimerkkien laskeminen laskurilla antaa konkreettisuutta ja auttaa ymmärtämään datan ja tunnuslukujen yhteyden. Hallinnan kannalta tämä on yhä varsin tärkeää. Tilastolliset analyysit suorittaa jokainen tutkija nykyisin omalla työpöydällään tietotekniikkaa hyödyntäen.

Ensimmäisessä laitoksessa erotettiin kuvaava ja päättelevä aines varsin kategorisesti toisistaan. Näitä on nyt pyritty yhdistämään vaikka teoksen rakenne on säilytetty. Uudemmat sovellukset (esim. SEM-tekniikat, monitasomallit yms.) on edelleen rajattu käsittelyn ulkopuolelle. Ne kuuluvat syventäviin tai post-graduate -opintoihin.

Materiaali ei ole sähköisen oppimisen mahdollisuuksia hyödyntävä. Sähköinen kanava toimii vain materiaalin levittämisen helpottajana.

Otamme mielellämme palautetta ja teemme sen perusteella korjauksia tarpeen mukaan.

Materiaalia saa käyttää vapaasti ei-kaupallisessa yliopistojen ja avoimen yliopiston opetuksessa.

Syyskuussa 2001

Erkki Komulainen (Erkki.Komulainen@Helsinki.Fi)
Kai Karma (Kai.Karma@Siba.Fi)

Versioon 2.2, (1.1.2002) on tehty pieniä korjauksia.

I Muuttujien välisten yhteyksien kuvaus

Käyttäytymistieteet, samoin kuin tiede yleensäkin, ovat tavallisesti varsin kiinnostuneita ilmiöiden välisistä yhteyksistä. Joku voi esimerkiksi tutkia jonkin opetusmenetelmän ja oppimistulosten välistä yhteyttä, joku toinen voi selvitellä erilaisten kysymysten tai väittämien välisiä yhteyksiä persoonallisuustestissä jne. Yhteyksistä voidaan tehdä johtopäätöksiä, jotka auttavat pääsemään tutkitulla alueella jälleen hiukan eteenpäin. Jos vaikkapa tietyn tyyppinen opetus tuottaa parhaat tulokset, ts. opetusmenetelmän ja oppimistulosten välillä on yhteyttä, voidaan tietyin edellytyksin varovasti olettaa, että kyseessä on syy-yhteys: hyvät tulokset johtuvat ko. menetelmän käytöstä. Jos taas tiettyihin persoonallisuustestin väittämiin eri henkilöt ovat taipuvaisia vastaamaan samansuuntaisesti, voidaan olettaa, että ne (osiot, väittämät) mittaavat suunnilleen samaa aluetta, esimerkiksi rehellisyyttä, uteliaisuutta, itseluottamusta jne. Tällaista samansuuntaisten kysymysten tai väittämien joukkoa voidaan pitää testissä omana alakokonaisuutenaan, jolloin siitä voidaan esim. laskea summapistemäärä, jota voidaan käyttää itsenäisenä mittanaan jne.

Jotta meillä olisi johdonmukainen esimerkkiaineisto, oletamme että joku on kerännyt kolmeltakymmeneltä henkilöltä (N=30) joitakin oleellisiksi katsomiaan tietoja, jotka on kerätty oheiseksi raakapistematriisiksi. Sukupuolella katsotaan olevan yhteyksiä aineistossa oleviin muihin muuttujiin, joten se on otettu matriisiin mukaan. Se on merkitty tavan mukaan ykkösellä ja nollalla, mutta itse asiassa mitkä tahansa kaksi toisistaan erottuvaa koodia kelpaisivat yhtä hyvin. Tämähän johtuu siitä, että sukupuoli on kvalitatiivinen, laadullinen muuttuja, eikä tällöin käytettyjen lukujenjärjestyksellä lukusuoralla ole väliä, ts. ne eivät kuvaa minkään ominaisuuden määrää. Viriketausta oletetaan mitatuksi asteikolla, joka perustuu tietoihin esim. kodissa olevien kirjojen määrästä, vanhempien harrastuksista, vanhempien ja lasten yhdessä viettämästä ajasta jne. Nämä tiedot on koottu kolmiportaiseksi koodiksi: huono, keskinkertainen ja hyvä. Verbaalinen (kielellinen) kyky on pyritty mittaamaan tähän tarkoitukseen tehdyllä testillä sekä järkeilykyky omallaan. Koulun päästötodistuksesta on otettu kieliaineiden keskiarvo, sekä matematiikan numero. Lopuksi on tavalla tai toisella, vaikkapa laskemalla tietty määrä tenttiarvosanoja yhteen, arvioitu ko. henkilöiden opintomenestystä koulun jälkeen:

	MUUTTUJAT						
	1	2	3	4	5	6	7
kh							
1	1	1	22	37	63	08	13
2	0	2	26	32	82	07	16
3	0	3	29	33	96	05	17
4	1	3	24	36	73	08	15
5	1	2	24	35	67	09	16
6	0	2	28	36	92	07	19
7	0	1	23	34	72	07	14
8	1	1	24	33	72	06	13
9	1	3	26	37	79	09	17
10	0	3	29	34	97	08	19
11	0	2	28	35	96	07	18
12	1	2	26	36	86	07	17
13	1	2	27	37	87	08	16
14	0	2	25	34	82	07	18
15	1	1	20	33	62	06	14
16	1	3	26	40	81	10	18
17	0	3	28	36	92	09	17
18	0	2	27	35	89	06	16
19	0	1	26	34	79	07	15
20	0	1	22	31	72	06	14
21	1	1	21	33	69	06	13
22	1	2	25	36	75	08	15
23	1	2	26	35	76	07	14
24	0	2	27	34	86	07	16
25	1	3	28	37	84	08	17
26	0	3	30	36	85	07	18
27	0	2	27	34	89	06	17
28	1	2	25	36	79	07	16
29	1	1	21	36	59	07	13
30	0	1	24	34	71	07	14

Muuttujaselitykset:

1. Sukupuoli (0=nainen, 1=mies)
2. Viriketausta (1 — 3)
3. Verbaalinen testi
4. Järkeilytesti
5. Kieliaineiden keskiarvo (ilman desimaalipilkkua)
6. Matematiikan numero
7. Opintomenestys

1. Kahden muuttujan välinen yhteys

a) Korrelaatio

Muuttujien välisten yhteyksien perustana on tavallisesti tieto yhteyksistä pareittain, siis aina kahden muuttujan välillä. Tavallisimpia tapoja ilmaista tämä on korrelaatiokerroin (Pearsonin tulomomenttikerroin). Aineistossa olevat muuttujien väliset yhteydet käyvät ilmi korrelaatiomatriisista, johon on kerätty kaikkien muuttujaparien väliset korrelaatiokertoimet. Edellä esitetystä havaintomatriisista saadaan seuraava korrelaatiomatriisi, jota on vielä täydennetty muuttujien (aritmeettisilla) keskiarvoilla (viiva-X) ja hajonnoilla (standardipoikkeamilla, s) myöhempien laskujen helpottamiseksi.

	1	2	3	4	5	6	7
1. Sukupuoli	1.000[1])						
2. Viriket.	-.044	1.000					
3. Verb. testi	-.449	.745	1.000				
4. Järk. testi	.464	.468	.246	1.000			
5. Kielten ka	-.548	.641	.909	.079	1.000		
6. Matem. nro	.338	.459	.143	.738	-.015	1.000	
7. Op. men.	-.387	.756	.812	.316	.826	.291	1.000
$\bar{X}$	0.50	1.97	25.47	34.97	79.73	7.23	15.83
S	0.51	0.76	2.57	1.83	10.39	1.10	1.84

Matriisista voidaan lukea monia aineiston luonnetta valaisevia seikkoja. Ensimmäisellä muuttujalla, sukupuolella, on selviä, vaikkakaan ei kovin voimakkaita yhteyksiä kaikkiin muihin muuttujiin paitsi virike- taustaan (r=-.04). Koska miehiä on tässä merkitty suuremmalla luvulla (1) kuin naisia, ovat miehet parempia niissä suorituksissa, joihin suku- puolen korrelaatio on positiivinen. Vastaavasti ovat naiset parempia tehtävissä, joiden korrelaatio ensimmäiseen muuttujaan on negatiivinen. Voimme siis todeta, että tässä aineistossa ovat järkeilytehtävät ja matematiikka olleet miehillä paremmat, kielitehtävät ja opintomenestys puolestaan naisilla.

Huomiota herättävän korkeita korrelaatioita on esim. kielten keskiarvon sekä verbaalisen testin (r=.91), opintomenestyksen ja kielten keskiarvon (r=.83) sekä

opintomenestyksen ja verbaalisen testin (r=.81) välillä. Sitä mistä nämä korrelaatiot johtuvat, ei näistä luvuista näe, mutta muuttujien laadun ja mahdollisen lisätiedon avulla voidaan tehdä tästä seikasta enemmän tai vähemmän oikeaan osuvia päätelmiä. On esimerkiksi johdonmukaista ja uskottavaa, vaikkakaan ei varmaa, että verbaalisen testin ja kielten keskiarvon välillä on voimakas yhteys siksi, että molemmat ovat mittoja suunnilleen samasta asiasta, kielellisestä kyvykkyydestä. Samoin olisi ymmärrettävää, että kielellisten mittojen korkea yhteys opintomenestykseen syntyisi ko. opintoalueen kielellisestä painottuneisuudesta. Voidaan olettaa, että sekä opittavan materiaalin ymmärtäminen että sen esittäminen selkeästi tentissä olisivat voimakkaasti kielellisestä kyvystä riippuvia. Joskus voi yhteys syntyä myös siksi, että toinen muuttuja on toisen syy. Voimme uskoa, että ainakin huomattava osa viriketaustan ja kielellisen kyvyn yhteydestä johtuu siitä, että virikkeet ovat vaikuttaneet kykyyn. Ilman lisäperusteita ovat tällaiset syysuhdepäätelmät kuitenkin hyvin vaarallisia. Viriketaustan ja kielellisen kyvyn yhteys voidaan aivan yhtä hyvin selittää olettamatta niiden välistä syy-yhteyttä. Voidaan ajatella, että tämä kyky on periytyvä ja esiintyy siten lapsissa virikkeistä .jokseenkin riippumatta. Jos nyt lahjakkaat vanhemmat sekä hankkivat paljon kirjoja, harrastavat monia asioita jne. että saavat lahjakkaita lapsia, esiintyvät virikkeet ja lasten lahjakkuus yhdessä ilman, että toinen olisi toisen syy. Pikemminkin niillä on yhteinen syy; lahjakkaat vanhemmat.

Matriisissa esiintyy myös joitakin varsin matalia korrelaatioita. Esimerkiksi kielten keskiarvolla ei ole juuri mitään tekemistä järkeilytestin (r=.08) eikä matematiikan numeron (r=-.02) kanssa. Niinpä siis vaikkapa lahjakas matemaatikko voi olla kielissä lahjakas tai lahjaton, matematiikan kyvyn perusteella ei voi arvata, onko henkilöllä kielellistä kykyä vai ei.

Niin käyttökelpoinen kuin korrelaatiokerroin usein onkin, sillä on joitakin heikkouksia, jotka saattavat haitata oikeiden tai tarkkojen johtopäätösten tekoa. Se ilmaisee yhteyden vain ylimalkaisesti, keskimäärin, eikä anna mahdollisuutta tarkempaan analyysiin siitä, millä tavoin yhteys muodostuu. Saattaa esimerkiksi olla, että yhteys johtuu pääasiassa muutamien harvojen yksilöiden saamista äärimmäisistä arvoista, jolloin loppujen kohdalla ei korrelaatiota ole. Samoin voi käydä niin, että vain jollakin kohdalla muuttujan aluetta, esimerkiksi pienien arvojen kohdalla, voidaan havaita yhteyttä toiseen muuttujaan. Esimerkiksi monen tutkimuksen perusteella näyttää siltä, että musikaalisuus ja älykkyys korreloivat

positiivisesti, mutta lähinnä vain matalilla älykkyystasoilla; älykäs ihminen voi yhtä hyvin olla musiikillisesti lahjakas kuin lahjatonkin.

b) Ristiintaulukointi

Ristiintaulukointi on konkreettisuutensa vuoksi varsin tavallinen tapa esittää kahden muuttujan välinen yhteys. Siihen sisältyy mahdollisuus kuvata määrinä ja %-lukuina asioita, joten se soveltuu myös esitystavaksi hyvin. Kuulijalta tai lukijalta ei oleteta laajoja tilastollisia esitietoja.

Joskus yhteyden luonne poikkeaa lineaarisesta. Muuttujat saattavat myös olla vain luokittelutason muuttujia. Tällöin voidaan kahden muuttujan yhteyttä tarkastella ristiintaulukoinnin avulla. Sen avulla voidaan korrelaatiokertoimenkin kautta saatua tietoa usein tarkentaa tai havainnollistaa.

Voimme tässä havainnollistaa tilannetta tarkastelemalla lähemmin esi- merkkiaineistomme viriketaustan ja kieliaineiden keskiarvon välistä yhteyttä. Korrelaatiomatriisista näemme, että yhteyttä selvästikin on, korrelaatio on .64, joka on jo suhteellisen selvä yhteys. Parempi viriketausta liittyy siis tässä aineistossa yleensä parempaan menestykseen kieliaineissa. Tämä tieto on kuitenkin vielä suhteellisen ylimalkainen ja sitä voidaan tarkentaa muilla keinoilla, joista tässä sovellamme ensimmäiseksi ristiin taulukointia.

Tavan mukaan nimitämme selittävää tekijää (tai sellaiseksi ajateltavissa olevaa muuttujaa) X:ksi ja sijoitamme sen vaakasuoraan niin, että arvot kasvavat vasemmalta oikealle. Tässä tapauksessahan se on viriketausta, joka saa arvot 1, 2 ja 3. Selitettävä tekijä, joka tässä on menestys kieliaineissa, merkitään Y:llä ja sijoitetaan taulukkoon pystysuoraan. On loogista, että sen arvot sijoitetaan siten, että ne kasvavat alhaalta ylös, jolloin taulukko on itse asiassa X, Y -koordinaatisto, jonka arvot ovat luokiteltuja, epäjatkuvia. Jos muuttujat ovat laadullisia, kvalitatiivisia, ei niiden järjestyksellä luonnollisestikaan ole väliä. Koska meidän esimerkissämme kieliaineiden keskiarvo -muuttuja saa hyvin monia eri arvoja, on se käytännöllisintä luokitella harvempiin kategorioihin. Voimme vaikkapa erottaa toisistaan huonot, keskinkertaiset ja hyvät arvosanat ja merkitä niitä koodein 1,2 ja 3. Tällöin saamme siis 3*3 -taulukon. Ensimmäisen koehen-

kilön viriketausta on huono (1) samoin kuin keskiarvokin (6.3), joten hänen kohdallaan tehdään merkintä vasemmalle ylös, jossa sekä X että Y saavat arvon yksi. Seuraavan koehenkilön arvot molemmilla muuttujilla ovat keskinkertaisia, joten hänen merkkinsä tulee keskimmäiseen ruutuun. Kun kaikki henkilöt on käyty läpi, saadaan seuraavan kaltainen ristiintaulukko:

		Viriketausta (X)			
		1	2	3	Yht.
Kieli-	1	3	0	0	3
aineiden	2	6	6	4	16
keskiarvo	3	0	7	4	11
(Y)					
	Yht.	9	13	8	30

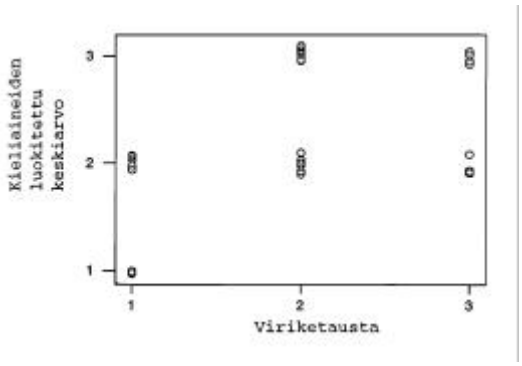
Tapauksia, joilla olisi huono viriketausta ja hyvä menestys kielissä (vasemmalla alhaalla) ei siis ollut lainkaan, sellaisia, joiden viriketausta on keskinkertainen ja menestys hyvä (keskellä alhaalla) on esiintynyt seitsemän kappaletta. Alhaalla ovat sarakkeiden summat, siis X-muuttujan jakauma sellaisenaan, ilman Y:tä ja vastaavasti oikealla rivisummat, jotka muodostavat Y-muuttujan jakauman.

Kieliaineiden keskiarvo-muuttujaa on luokitettu siten, että arvot alimmista 6.49 (kun desimaali on laitettu paikalleen) muodostavat luokan 1, arvot välillä 6.50 ...8.49 luokan 2 ja arvot 8.50 aina suurimpaan arvoon luokan 3. Ristiintaulukon tyypillinen tietokonetulostus järjestää luokkien koodiarvot kasvavaan järjestykseen taulukon vasemmasta yläkulmasta alkaen oikealle ja alas. Asioiden esittämiseksi joudutaan tietokonetulostusta lähes aina muokkaamaan parempaan esitysmuotoon. Taulukosta voidaan laskea rivi-, sarake- ja kokonaisprosenttilukuja. Taulukon käyttöä helpottaa suuresti myös solukohtaisten sattumalta odotettavissa olevien frekvenssien laskeminen. Näistä on tietoa myöhemmin kirjassa. Laske taulukosta myöhemmässä vaiheessa khiin neliö ja arvioi, onko viriketaustan ja kieliaineiden luokitetun muuttujan välinen yhteys sattumayhteydestä merkitsevästi poikkeava.

Taulukkoa tarkasteltaessa voidaan selvästi havaita, että suuria X:n arvoja pyrkivät seuraamaan suuret arvot Y:llä ja vastaavasti pyrkivät pienet arvot kummallakin muuttujalla esiintymään yhdessä. Tämähän on tieto, joka sisältyi jo suhteellisen korkeaan positiiviseen korrelaatioon eikä siis ole mitään varsinaisesti uutta. Tarkempi tarkastelu osoittaa kuitenkin myös sellaista, mitä korrelaation perusteella ei olisi voinut tietää: vain huonoimmalla viriketaustalla on

yhteyttä huonoon menestykseen. Ne, joiden tausta on keskinkertainen, ovat menestyneet jokseenkin samalla tavoin kuin nekin, joilla on hyväksi arvioitu tausta. Taustan ja menestyksen välillä ei olekaan täysin suoraviivaista, lineaarista yhteyttä, jossa tietyn suuruista kasvua toisella muuttujalla edustaa koko ajan vastaava (suhteellinen, keskimääräinen) kasvu toisella. Yhteys on käyräviivainen, kurvilineaarinen, siten että X:n kasvua ensin vastaa selvä kasvua myös Y:llä, mutta X:n kasvaessa edelleen ei Y enää muutukaan (ainakaan samassa suhteessa).

Jo nyt voi pohtia sitä kuinka kieliaineiden rivin 3 tapaukset (F=11) jakautuisivat oman rivinsä soluihin, jos jakauma noudattaisikin viriketaustan erittelemätöntä kokonaisjakaumaa eli alinta yhteensä riviä. Samoin voi pohtia kuinka viriketausta arvon 1 sarakkeen 9 tapausta jakautuisivat sarakkeensa soluihin, jos ne noudattaisivatkin kieliaineiden erittelemätöntä kokonaisjakaumaa eli viimeinen yhteensä sarake. Kun tuon oivaltaa, on käsittänyt mitä tarkoitetaan riippumattomuusluvuilla eli sattumalta odotettavissa olevilta frekvensseiltä. Ne ovat teoreettisia (odotus-) arvoja ja sen vuoksi ne ilmoitetaan vaikkapa kahden desimaalin tarkkuudella. Katso myös lukua d) kontingenssikerroin.



Ristiintaulukko on muutettu alkeelliseen graafiseen muotoon. Y-muuttujan suunnassa kuhunkin arvoon on lisätty pieni satunnaistekijä ("tärinä"), jotta arvot erottuisivat toisistaan. Käyräviivainen yhteys havainnollistuu. Huomaa, että graafiset esitykset laaditaan useimmiten näin: esityksen molempien muuttujien pienin arvo sijoittuu vasemmalle alas. Taulukossa on 30 X-Y-pistettä.

Myös ristiintaulukoinnin ongelmia alkaa tulla esille. Alkuperäisiä muuttujan arvoja täytyy karkeistaa (uudelleen luokitella). Alkuperäinen ja luokitettu muut-

tuja eivät sisällä täysin samaa informaatiota. Luokitettaessa informaatiota yleensä kadotetaan. Tässä alkuperäisen ja luokitetun kieliaineiden keskiarvo-muuttujan korrelaation on kuitenkin niinkin korkea kuin r=.898.

Toinen ilmeinen ongelma on, että jo 3 kertaa 3 taulukon solukkoon 30 tapausta on aika niukasti. Kun oikeassa tutkimuksessa yleensä jatketaan tarkastelemalla asiaa erikseen vaikkapa sukupuolen mukaan, alkaa suurikin aineisto pian tuntua pieneltä solumäärän kasvaessa. Minimiksi mainitaankin usein, että pienin solu-frekvenssi pitäisi olla vähintään 5 tai pienin khiin neliön odotusarvo saa olla (havaitusta arvosta riippumatta) 5.

Kolmantena on mainittava vaikkapa se, että ei ole mitään sääntöä miten (tasa-välisesti ymv. tavalla) luokitus tehdään. Jos alkuperäinen jakauma on kovin vino, on vaikea käyttää yhtä suuria luokkavälejä.

Perustiedon muistamisen kokeilemiseksi selvitä itsellesi: mitkä ovatkaan yllä suoritetun luokituksen luokkakeskukset, luokkarajat ja luokkaväli (primaariar-voina ilmoitettuina)?

Yllä olevasta graafista saisi nykyvälinein helposti myös kolmiulotteisen kuva-uksen, jossa solun frekvenssit nousevat pylväinä. Sellaiset kuvaukset puoltavat paikkaansa huolellisissa esityksissä. Täytyy muistaa kuitenkin, että sanoma-lehtikirjoituksella ja tieteellisellä tekstillä on omat kirjoittamattomat sääntönsä grafiikan käytössä. Graafisten esitysten käytöstä tieteellisessä tekstissä löytyy omia erityisteoksiaan ja oppaita.

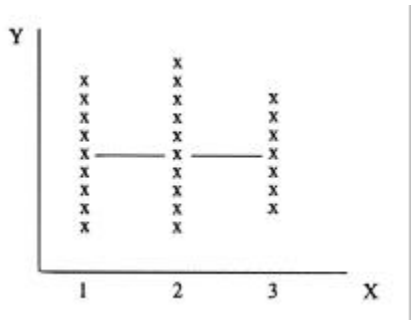
c) Eta-kerroin

Käyräviivaiset yhteydet ovat tutkimuksessa aina hieman ongelmallisia. Korre-laatiokerroin toimii ikään kuin yhteys olisi lineaarinen: jos yhteys on selvästi non-lineaarinen, ei korrelaatiokerroin enää ole hyvä indikaattori, ja sen arvot jäävät todellista yhteyttä pienemmiksi. Muita kertoimia on kuitenkin usein hankala käyttää; esimerkiksi useamman muuttujan välisiä yhteyksiä kuvaavat monimuuttujamenetelmät perustuvat yleensä korrelaatiomatriisiin. Jos tarkoitus on kuvata vain kahden muuttujan välistä yhteyttä, jonka epäillään olevan käy-räviivainen, voi korrelaatiokertoimen sijasta tietyin edellytyksin laskea korre-

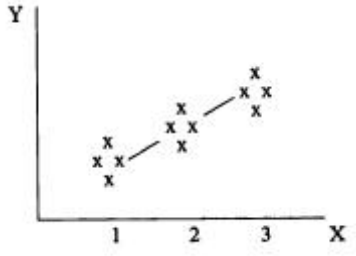
laatiosuhteen eli eta-kertoimen. kuten sitä kreikkalaisen symbolinsa mukaan usein nimitetään.

Korrelaatiosuhteen ymmärtämiseksi on syytä ensin paneutua vaihtelun osiin jakamisen ideaan, joka on tärkeä periaate monissa kehittyneemmissä myöhemmin esitettävissä tilastollisissa menetelmissä. Muuttujissa esiintyvän vaihtelun voi jakaa osiin useallakin tavalla, mutta tässä yhteydessä olennainen on jako ryhmien sisäiseen ja ryhmien väliseen vaihteluun. Tällöin on kyseessä juuri sen kaltainen tilanne, joka meillä on omassa esimerkissämme. Selittävä muuttuja X (tässä: viriketausta) on epäjatkuva tai luokiteltu, selitettävä tekijä Y (tässä: kieliaineiden keskiarvo) on joko jatkuva tai epäjatkuva, kuitenkin asteikkotyypiltään ainakin lähellä intervallitasoa, niin että siitä voidaan mielekkäästi laskea keskiarvoja ja hajontoja.

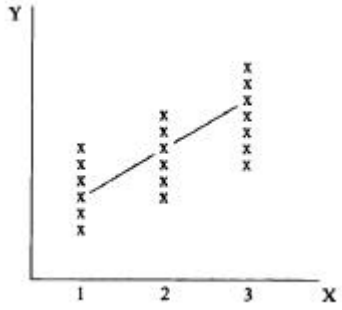
Tällaisessa tapauksessa voidaan esiintyvän vaihtelun jakaantumista eri tekijöille ehkä aluksi parhaiten tarkastella ääritapauksia kuvaavien esimerkkien avulla. Kuvitellaan ensin, että aineisto X-Y -koordinaatistoon vietynä näyttäisi seuraavalta:



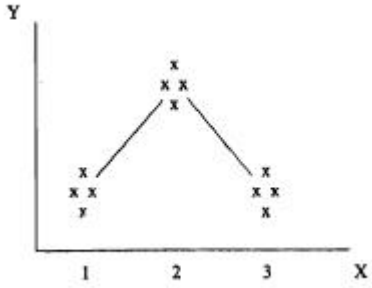
Kunkin X-arvon kohdalla olevassa ryhmässä on selvää vaihtelua: kussakin ryhmässä on sekä huonoja että hyviä Y-arvoja. Ryhmien keskiarvot ovat kuitenkin Y-muuttujalla aivan samat. Ryhmien tasolla tarkasteltuna vaihtelua ei siis ole. Voidaan sanoa, että kokonaisvaihtelu koostuu pelkästään ryhmien sisäisestä vaihtelusta, ryhmien välistä vaihtelua ei ole. Toinen ääritapaus olisi seuraava:



Nyt on tilanne päinvastainen: ryhmien sisällä kaikki saavat saman arvon (niin tarkoin kuin sen voi piirtää) kun taas ryhmien välillä on huomattavia eroja. Nyt sanottaisiin, että kokonaisvaihtelu koostuu pelkästä ryhmien välisestä vaihtelusta ja ryhmien sisäistä vaihtelua ei ole. Mielenkiintoisin ja todellisuudessa yleensä esiintyvä tapaus olisi seuraavankaltainen:



Tässä tapauksessa on vaihtelua sekä ryhmien sisällä että niiden välillä. Kokonaisvaihtelun voidaan sanoa koostuvan kahden komponentin, ryhmien sisäisen ja niiden välisen vaihtelun summasta. Eta-kerroin muodostuu ryhmien välisen ja kokonaisvaihtelun suhteesta. Jos vaihtelua on vain ryhmien välillä, saa kerroin arvon yksi, kun taas tapauksessa, jossa kaikki vaihtelu on ryhmien sisäistä, on eta:n arvo nolla. Kun tätä ajatusta tarkastelee lähemmin, huomaa että kerroin ei ole riippuvainen yhteyden muodosta. Ryhmien sisäinen vaihtelu on nolla, kun kaikilla ryhmän jäsenet ovat lähes samassa Y:n arvossa. Tämä ei liity mitenkään siihen minkä koodin tai arvon ryhmänjäsenet X-muuttujalla saavat. Voimme vaihtaa ryhmien paikat X-akselilla ilman että Y-muuttujan kokonaisvarianssi muuttuu tai että ryhmien sisäisten varianssien määrät muuttuisivat. X-muuttuja ymmärretään siis siten, että sen arvoja käsitellään laadullisen muuttujan tapaan eikä koodiin liity mitään kvantitatiivista ajatusta. Voimme tarkastella esimerkkinä seuraavaa tapausta:



Yhteyden voimakkaasta käyräviivaisuudesta johtuen on korrelaatio nolla. Siitä huolimatta koostuu kokonaisvaihtelu ainoastaan ryhmien välisestä vaihtelusta ja eta-kerroin saa arvon yksi. Kerrointen tasolla näkyy siis käyräviivaisuus siinä. että eta saa korrelaatiota korkeamman arvon. Jos yhteys on täysin lineaarinen, on korrelaatiokerroin identtinen etan kanssa. Eta-kertoimen laskemiseksi meidän on siis hankittava mitat kokonaisvaihtelusta sekä ryhmien välisestä vaihtelusta. Tähän tapaukseen sopiva vaihtelun mitta on neliösumma (SS, sum of squares). Neliösumman voisi hieman kiertäen määritellä "varianssiksi, ennen kuin se on jaettu numeruksella", ts. se on varianssin kaavassa osoittajana. Jakajana on parempi käyttää kuitenkin vapausasteita (df=N-1).

$$SS = \sum (X - \overline{X})^2 \qquad SS = (N - 1) \cdot s^2$$

Omasta esimerkistämme meillä on jo tiedossa y-muuttujan kokonaisvaihtelu: aiemmin esitetyssä korrelaatiomatriisissahan oli mainittu kieliaineiden hajonnaksi 10.39. Kun tämä korotetaan toiseen, saadaan varianssi 107.95. Neliösumma on tämä kerrottuna tapausten lukumäärällä (tarkemmin ottaen vapausasteilla df=N-1) 29*107.95 = 3131.

Menemättä laskutoimitukseen tarkemmin voidaan todeta, että ryhmittäiset varianssit ovat seuraavat: heikoimman viriketaustan ryhmässä oli kieliaineiden keskiarvo-muuttujan varianssi 39.44, keskiryhmässä 61.61 ja parhaassa 72.13. Kun nämä kerrotaan ryhmien tapausten määrillä (tark. df:llä), saadaan neliösummat 9*39.44=315, 12*61.61=739 ja 7*72.13=505. Nämä siis laskettiin kustakin ryhmästä erikseen eli ne ovat ryhmien sisäisiä neliösummia. Niiden hieman pyöristetty summa on 1560. Se on sisäisen vaihtelun mitta. Koska se on vain noin puolet y-muuttujan kokonaisvaihtelusta (3131), voimme mielessämme todeta, että aineistossa täytyy olla myös huomattavaa ryhmien välistä vaihtelua.

Korrelaatiosuhdetta vartenhan tarvittiin ryhmien välisen vaihtelun mitta, jota meillä ei vielä ole. Koska vaihtelu voi olla vain ryhmien sisäistä tai niiden välistä, saadaan ryhmien välinen vaihtelu vähentämällä kokonaisvaihtelusta ryhmien sisäinen vaihtelu: SSb=SSt-SSw (SSbetween=SStotal - SSwithin), tässä tapauksessa SSb=3131 - 1560=1571.

Se osuus, jonka X tilastollisesti selittää Y:stä, saadaan jakamalla ryhmien välinen vaihtelu kokonaisvaihtelulla, siis:

$$\eta^2 = \frac{SS_b}{SS_t}$$

Tämä .52 on eta-kertoimen neliö.

Voidaan siis sanoa, että tuntemalla X voidaan Y:n vaihtelusta tietää 52 %. Tämä on aivan vastaava asia kuin korrelaatiokertoimen neliö: korottamalla kerroin toiseen saadaan tietää, kuinka paljon toinen muuttuja selittää (tilastollisesti ottaen, lineaarisen regression kautta) toisesta. Jos esim. korrelaatio on .80, selittävät muuttujat toisistaan .802 = .64 eli 64 %. Toisin päin saadaan selitysosuudesta alkuperäinen kerroin ottamalla siitä neliöjuuri. Niinpä äsken lasketusta 52 %:sta tulee eta-kerroin = .72. Perinteisesti korrelaatiota käytetään sellaisenaan. Eta ilmaistaan yleensä vain eta-toiseen -kertoimena.

Nyt meillä on tieto siitä, kuinka paljon X:llä voidaan selittää Y:tä, kun korrelaation lineaarisuusoletus ei ole mukana haittaamassa yhteyden tarkastelua. Samoin meillä on vahvistus sille, että yhteys todella on käyräviivainen: korrelaatiokerroinhan oli .64, minkä eta-kerroin selvästi ylittää.

Tässä asia on laskettu eräällä tilasto-ohjelmalla ja siihen liittyvällä alkeellisella graafisella kuvauksella:

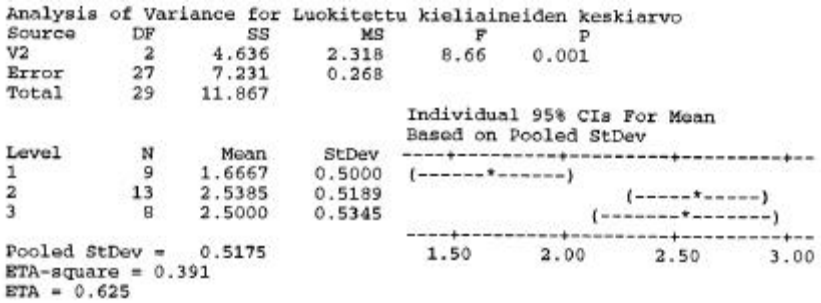
```
Analysis of Variance for Kieliaineiden keskiarvo
Source     DF       SS        MS        F        P
Viriketa    2    1570.2     785.1    13.59    0.000
Error      27    1559.7      57.8
Total      29    3129.9

                                Individual 95% CIs For Mean
                                Based on Pooled StDev
Level      N      Mean     StDev  -+---------+---------+---------+-----
1          9    68.778     6.280  (------*-----)
2         13    83.538     7.849                    (----*-----)
3          8    85.875     8.493                     (------*------)
                                  -+---------+---------+---------+-----
Pooled StDev =    7.600          64.0      72.0      80.0      88.0
Eta-squared = 0.502
Eta = .708
```

Taskulaskimella pääsi siis aivan riittävään tarkkuuteen. Erot johtuvat pyöristyksestä. Taskulaskimella laskien on syytä huomata myös ero hajonnoissa s (jakajana N-1 eli vapausaste) ja S (vanhemmissa kirjoissa, jakajana N).

Palaa tähän esimerkkiin myöhemmin ja tutki sitä, kun yksisuuntainen varianssianalyysi ja siihen liittyvä hypoteesin testaus F-jakaumineen on tullut tutuksi. Myös kuviossa käytettyjen luottamusrajojen laskeminen ja käyttö tulee tutuksi myöhemmin. Jos numerot kiinnostavat, niin voit selvittää itsellesi, mikä on Pooled StDev hajontana ja miten se on laskettu! Samoin vapausasteet (df) ja keskineliö (MS) eivät ole vielä tuttuja asioita, mutta niihinkin tutustutaan.

Jos eta-(toiseen) -kerroin ja siihen liittyvät laskelmat suoritettaisiin ristiintaulukoinnissa käytetyillä karkeammilla arvoilla (1,2 ja 3), saataisiin seuraava tulos:



Huomaat, että numeeriset arvot muuttuvat, kun y -muuttujasta käytetään luokitettuja arvoja. Johtopäätösten osalta tulos pysyy samana. Eta-kertoimeen ja yksisuuntaiseen varianssianalyysiin käytetään säännönmukaisesti y-muuttujan mahdollisimman tarkkoja (raakapistemäärä)arvoja. Karkeistaminen luokittelemalla vie tehoa pois tilastollisista tarkasteluista, mikä seikka tässäkin näkyy F-suhteen pienenemisenä ja selitysosuuden madaltumisena.

Vaikka eta-toiseen -kerroin on joskus havainnollinen ja käyttökelpoinen esim. käyräviivaisten yhteyksien löytämisessä, on sillä haittoja, jotka ovat tehneet sen

yksinään käytettynä melko harvinaiseksi. Ensinnäkin tämä kerroin on epäsymmetrinen: X:n korrelaatiosuhde Y:hyn ei ole välttämättä sama kuin Y:n suhde X:ään. Toiseksi, ei ole juuri tilastollisia menetelmiä, jotka käytännössä pohjautuisivat eta-kertoimeen sillä tavoin kuin monet menetelmät pohjautuvat korrelaatioon. Tästä kertoimesta on siis vaikea päästä eteenpäin. Kolmanneksi, etan suuruus riippuu siitä, miten X-muuttujan luokitus tehdään. Luonnollisin on tapaus, jossa X-muuttuja on jo valmiiksi ryhmiteltyaineiston luonnetta vastaaviin tasoihin eikä tätä keinotekoisesti muuteta. Jos X luokitellaan, on se tehtävä niin, että kuhunkin luokkaan jää kohtuullinen osa tapauksia. Jos korrelaatiota ja etaa halutaan verrata keskenään, on ne laskettava samalla tavoin luokitellusta aineistosta.

Edellä esitetty koski määrällisten, kvantitatiivisten muuttujien välistä yhteyttä. Jos muuttujat ovat kvalitatiivisia, laadullisia, ei korrelaatiota tai korrelaatiosuhdetta (eta) voida yleensä laskea tai käyttää. Eihän ole mielekästä laskea hajontoja tai puhua siitä, kuinka toinen muuttuja kasvaa toisen kasvaessa, jos mistään todellisesta määrästä ei ole kysymys. Täysin laadullistenkin muuttujien yhteydestä voidaan tietyssä mielessä kuitenkin puhua ja käymme nyt tarkastelemaan sitä lähemmin.

Kun tehdään varianssianalyyseja joissa on useita selittäviä muuttujia yhtä aikaa käytössä y:n tilastollisessa tarkastelussa, kunkin selittäjän itsenäinen selitysosuus (kun muut tekijät on tilastollisesti otettu huomioon, vakioitu) ilmaistaan usein eta-toiseen nimisenä tunnuslukuna. Se rinnastuu myöhemmin regressioanalyysin yhteydessä esitettävään omaosuuteen eli semipartiaalikorrelaation neliöön. Mittalukujen kanssa on kuitenkin syytä olla varovainen. Spss:n partial eta-squared lasketaan poikkeavalla tavalla (moduli GLM/Univariate)!

d) Kontingessikerroin

Kuvitellaanpa että jonkin tuotteen, vaikkapa kahvin, valmistaja on kiinnostunut siitä, jakaantuuko eri laatujen käyttö ostajan asuinpaikan mukaan. Olkoot kahvilaadut A, B ja C ja asuinpaikat luokiteltu kaupunkeihin, pieniin taajamiin sekä maaseutuun. Pienen haastattelukierroksen jälkeen voimme saada vaikka seuraavan ristiintaulukon:

Asuinpaikka	Valittu laatu A	B	C	
kaup.	20	10	20	50
taaj.	10	20	10	40
maas.	5	5	20	30
	35	35	50	120 = N

Siis kaksikymmentä kaupunkilaista oli pitänyt eniten kahvista A, kymmenen kaupunkilaista kahvista B jne. Onko nyt asuinpaikalla ja kahvin valinnalla jotakin yhteyttä ja jos on, mistä sen voi nähdä? Tähän kysymykseen voidaan saada vastaus katsomalla ensin tietoa siitä, miten valinnat jakaantuisivat, jos mitään yhteyttä ei olisi. Toisin sanoen: kun kerran tiedämme, että kaupunkilaisia on 50, taajamissa asuvia 40 ja maalaisia 30 ja samoin tiedämme, että kahvit valittiin (jos asuinpaikka ei vaikuta) suhteissa 35/35/50, mitkä arvot pitäisi taulukossa olla, jos yhteyttä asuinpaikan ja valinnan välillä ei olisi?

Hieman teknisemmin tämä voidaan sanoa siten, että meidän on laskettava/arvioitava kuhunkin ruutuun odotetut frekvenssit fe (expected frequency) siellä jo olevien havaittujen frekvenssien fo (observed frequency) lisäksi oletuksella, että yhteyttä muuttujien välillä ei ole. Käytännössä tämä tapahtuu kertomalla kutakin ruutua vastaava sarakesumma ja rivisumma keskenään ja jakamalla tulo numeruksella:

$$f_e = \frac{f_r \cdot f_k}{N}$$

Meidän esimerkissämme olisi vasemmalla ylhäällä olevan ruudun odotusarvo siis 50*35:120 = 14.6. Kun muut on laskettu vastaavasti, voidaan taulukko täydentää odotusarvoilla seuraavasti:

		A	B	C	
K	fo	20	10	20	50
	fe	14.58	14.58	20.83	
T	fo	10	20	10	40
	fe	11.67	11.67	16.67	
M	fo	5	5	20	30
	fe	8.75	8.75	12.50	
		35	35	50	120

Asia liittyy todennäköisyyslaskentaan. Aineiston mukaan tn olla kaupunkilainen on 50/120 eli p=0.4167. Samoin aineistosta saatu tn käyttää kahvilaatua A on 35/120 eli p=0.2917. Jos asumismuoto ja kahvilaadun käyttö olisivat toisistaan riippumattomia olisi tn olla kaupunkilainen ja A:n käyttäjä näiden erillisten todennäköisyyksien tulo eli p=0.1215. Tällaisia odotettaisiin olevan siis 0.1215 koko määrästä N (siis 0.1215*120 = 14.58). Odotus- frekvenssit ovat siis teoreettisia tapausmääriä 120:stä, jotka solussa olisivat, mikäli muuttujat olisivat toisistaan riippumattomia. Teoreettiset tapausmäärät ilmaistaan desimaaliosaa käyttäen.

Khiin neliö lasketaan aina frekvensseistä. Taulukon kuvaus ja tuloksen ymmärtäminen sujuvat monilta paremmin %-lukujen avulla. Odotusarvoprosentin käsite on kuitenkin tilastoterminologialle vieras. Voimme kuitenkin vallan hyvin sanoa, että kaupunkilaisista 40 % käyttää kahvilaatua A kun sattumaodotus sille on 29 %. Tai: kahvilaatua C käyttää aineistossa 42 % tutkituista kun taas maalla asuvista sitä käyttää peräti 67 %. Ei ole ainoita oikeita tapoja taulukon kuvaamiseen. Kuvauksen pitää olla koherentti (yhtäpitävä) taulukosta saatujen tilastosuureiden kanssa. Osa kuvauksista, jotka joskus tuntuvat hyviltä (tai ovat toiveiden mukaisia) voi siis olla ihan vääriä, jos ei ole tarkkana. Tekstin ja taulukon täytyy antaa asioista samantapainen kuva kuitenkaan samaa toistamatta.

Taulukkoa tarkastelemalla voimme heti todeta, että odotusarvot poikkeavat havaituista ainakin jonkin verran. Esimerkiksi kaupunkilaiset ovat pitäneet kahvista A enemmän kuin heidän määränsä olisi antanut aiheen olettaa, taajamissa asuvat kahvista C vähemmän kuin heille sattumanvaraisessa jaossa tulisi jne. Jotta saisimme tästä yhteydestä määrällisen indikaattorin, meidän täytyy ensin

laskea khiin neliön nimellä tunnettu tunnusluku. Se perustuu havaittujen ja odotettujen frekvenssien eroihin ja on kaavan muodossa seuraavanlainen:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Laskemme ensin jokaista ruutua vastaavan havaitun arvon ja odotusarvon erotuksen, korotamme nämä toiseen, jaamme odotusarvolla ja lopuksi laskemme kaikki näin saadut solukohtaiset luvut yhteen. Ensimmäistä ruutua vastaava arvo on siis (20-14.6)2:14.6 = 2.00, muut saamme vastaavasti. Lopullisesta summamerkin osoittamasta yhteenlaskusta tulee tulos 20.05. Solukohtaisesti:

$$\begin{aligned} x^2 = {} & 2.01+1.44+0.03+ \\ & 0.24+5.94+2.67+ \\ & 1.61+1.61+4.50 = 20.05 \end{aligned}$$

Khiin neliön arvo on siis 20.05, mutta tämä ei sellaisenaan ole muuttujien välisen yhteyden indikaattori, vaan yhteyden tilastollisen merkitsevyyden tunnusluku, joka käsitellään vasta myöhemmin. Tämän arvon voimme kuitenkin helposti muuttaa yhteyden voimakkuutta kuvaavaksi luvuksi, kontingenssikertoimeksi, seuraavalla tavalla:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Meidän esimerkissämme kontingenssikertoimen arvoksi tulee .38.

Vaikkakaan kontingenssikerroin ei ole sama korrelaatiokertoimen kanssa, voidaan sitä tulkita karkeasti samaan tapaan. Saatu yhteys, .38 on siis heikko, mutta kuitenkin varteenotettava. Ilmeisesti eri asuinpaikoissa suositaan ko. kahvilaatuja hieman eri tavoin. Kerroin on aina positiivinen, eikä etumerkki olisikaan mielekäs täysin kvalitatiivista aineistoa käsiteltäessä.

Khiin neliö perustuu siis kustakin solusta syntyvään elementtiin ja niiden summaan. Khiin neliön jakaumasta voimme päätellä, voimmeko hyväksyä vai hylätä sen tilastollisen hypoteesin, että havaittu yhteys on sattumavaihtelun rajoissa. Esimerkissämme tällainen nollahypoteesi voidaan hylätä hyvin pienellä riskillä olla väärässä (Khii-toiseen=20.05, df=4, p<.001). Yksittäisien solujenkin kohdalla voidaan varovasti päätellä missä kohtaa taulukkoa havaittu frekvenssi poikkeaa merkitsevästi odotusarvosta (eli missä havaittu prosentti ja odotusprosentti eroavat toisistaan). Ottamalla positiivinen neliöjuuri kustakin khiin neliön elementistä, merkitsevänä voidaan pitää soluja, joissa tuo arvo on suurempi kuin kaksi. Mutta varauksin: tällainen post-hoc -tarkastelu tuottaa tunnetusti enemmän merkitsevyyksiä kuin käytetty riskitaso sallisi. Suuntaa antavana sitä voidaan pitää. Kyseisellä tavalla laskettu solun elementin arvo kulkee nimellä standardoitu residuaali.

Nelikenttäkorrelaatio 2*2 -taulukosta (phi) saadaan Khiin neliön avulla. Khiin neliö jaettuna N:llä ja tästä luvusta neliöjuuri.

2. Ennustaminen muuttujalta toiselle

Varsin usein joudumme tekemään päätelmiä yhdestä muuttujasta toisen muuttujan tunnettujen arvojen perusteella. Voimme kysyä, mikä on henkilön todennäköinen opintomenestys, kun hänen koulumenestyksensä tunnetaan; mikä on paras arvio aikuisiän aktiivisuudesta, kun lapsuudenaikainen virikeympäristö tunnetaan jne. Myös arvioiden tekeminen yhdeltä muuttujalta toiselle samanaikaiselle muuttujalle on ennustamista, prediktiota, tässä hieman arkikieltä laajemmassa merkityksessä. Niinpä voimme vaikkapa "ennustaa" menestyksen yhdessä testissä, kun suoritus toisessa tunnetaan.

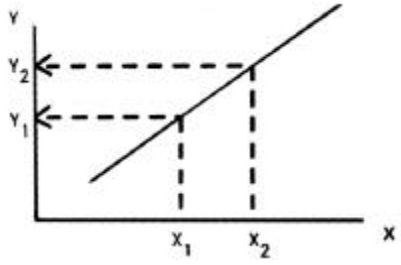
Jos pitäydymme muuttujien suhteen hyvin yksinkertaisessa jaottelussa kvantitatiivinen vs. kvalitatiivinen, voimme laatia nelikentän muotoon tehdyn typologian ennustamistilanteesta muuttujalta toiselle. Kun sekä ennustemuuttujia että kriteerimuuttujia voi olla yhtä aikaa analyysissä useita (monimuuttujaiset tarkastelut) ja kun sekä kvalitatiivisia että kvantitatiivisia voi olla niitäkin yhtä aikaa mukana joko x-muuttujissa, y-muuttujissa tai molemmissa, voidaan ymmärtää, että asia ei ole mikään kovin yksinkertainen.

kvalitatiivinen x, kvalitatiivinen y	kvalitatiivinen x, kvantitatiivinen y
kvantitatiivinen x, kvalitatiivinen y	kvantitatiivinen x, kvantitatiivinen y

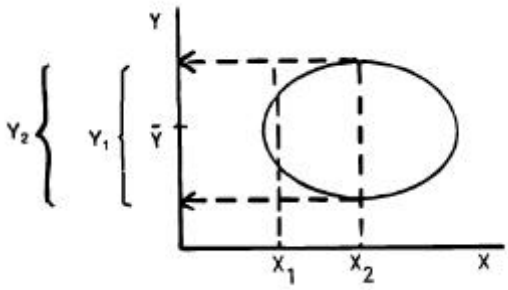
Käsittelemme nyt typologian ruutua, jossa kummatkin muuttujat ovat kvantitatiivisia. Ennusteita (prediktori, X) yksi ja kohdemuuttujia (kriteeri, Y) yksi. Tästä asia laajenee tilanteeseen, jossa ennustemuuttujia voi olla useita, mutta kriteerimuuttujia on edelleen vain yksi. Korrelaatiosta etenemme siis yhteiskorrelaatioon, yksinkertaisesta regressiosta multippeliregressioon. Selityksen kohteena voi olla useita muuttujia yhtä aikaa. Tällöin puhutaan kanonisesta analyysista, MANOVA:sta (Multivariate Analysis of Variance) tai yleisestä lineaarisesta mallista. Useiden Y -muuttujien tilanne on kuitenkin rajattu tämän esityksen ulkopuolelle.

Jotta prediktiota voisi tehdä, on muuttujien välinen yhteys tunnettava. Tavallisesti tämä yhteys merkitsee käytännössä korrelaatiota. Jotta siis voisimme vaikkapa edellisen esimerkin mukaisesti tehdä arvion jonkun henkilön testimenestyksestä toisen testin perusteella, on meillä oltava jossakin vaiheessa aineisto, jolle on tehty molemmat testit. Tästä aineistosta hankittu korrelaatiota käytetään myös mahdollisesti myöhemmin hankitun laajemman toisen aineiston ja siellä suoritetun ennustamisen pohjana (ristiinvalidointi).

Korrelaatio merkitsee eräässä mielessä juuri mahdollisuutta sanoa jotakin puhdasta arvausta parempaa arviota toisesta muuttujasta toisen perusteella. Mitä suurempi korrelaatio kahden muuttujan välillä on, sitä pitävämpiä arvioita voimme tehdä. Voimme lähteä tarkastelemaan tätäkin ongelmaa lähemmin ääriesimerkkien avulla. Kun kahden muuttujan arvoja kullakin tilastollisella yksiköllä (usein koehenkilö) kuvaavat pisteet piirretään suorakulmaiseen koordinaatistoon, saadaan korrelaation graafinen esitys, korrelaatiotaulu. Täydellisessä korrelaatiossa, jonka itseisarvo on siis yksi, asettuvat pisteet samalle suoralle. Tällaisessa tapauksessa vastaa aina kutakin X-arvoa vain yksi täsmälleen määriteltävissä oleva Y-arvo, ts. prediktio on täydellistä, ennusteet varmoja:

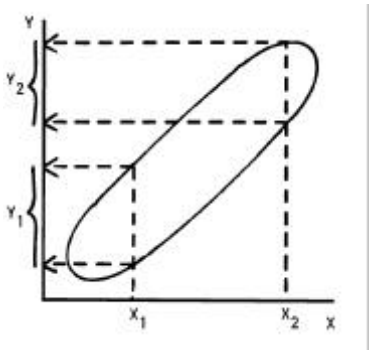


Toisessa ääritapauksessa, nollakorrelaation vallitessa muuttujien välillä, ei toisesta muuttujasta ole itse asiassa mitään apua ennusteen teossa. Tässä tapauksessa asettuvat pisteet joukoksi, jonka kumpikaan pää ei ole toista korkeammalla. Kun nyt valitsemme minkä tahansa X-arvon ja haemme vastaavia arvoja Y:ltä, päädymme aina tietyn kokoiselle alueelle, joka on symmetrisesti Y:n keskiarvon (Y) molemmin puolin:



Kun X-muuttujasta ei ole apua ennusteen teossa, parasta mitä voimme tehdä on valita Y -muuttujalta sen todennäköisin arvo eli sen keskiarvo. Tämä keskiarvo on siis "paras arvaus" silloin kun X:n ja Y:n välillä on nollakorrelaatio.

Tavallisin tapaus on jälleen se, Jossa korrelaatiolla on jokin nollasta poikkeava, itseisarvoltaan ykköstä pienempi arvo. Tällöin graafisessa kuvauksessa tulevat pisteet viistoon, pitkänomaiseen joukkoon, joka on sitä kapeampi ellipsi mitä suurempi korrelaatio on. Kun nyt haemme erilaisia X-arvoja vastaavia Y-arvoja, huomaamme, että ne eivät ole täsmällisiä, vaan sijaitsevat tietynkokoisella alueella, mutta ne ovat silti pelkkää arvausta parempia. Kunkin todellisen Y:n arvon poikkeama ennusteesta (suora) on pienempi kuin nollakorrelaation tilanteessa:



Itse asiassa me emme päädy Y-muuttujalla sellaiseen tarkkarajaiseen alueeseen, jollaiselta ne kuvassa näyttävät, vaan jakaumaan, jonka eri kohdissa Y-arvot sijaitsevat tietyllä todennäköisyydellä, mutta pääperiaate on aivan sama. "Paras arvaus" sijaitsee tämän todennäköisen alueen keskiosasta hieman koko jakauman keskiarvoa kohti, koska siellä tapausten frekvenssi on suurempi, ne ovat todennäköisempiä. Tärkeintä on huomata se, että epävarmuus pienenee korrelaation kasvaessa, siis pisteiden muodostaman kuvion tullessa kapeammaksi.

Yhteenvetona voimme todeta, että tehtäessä ennustetta muuttujalta toiselle on muuttujien välisellä korrelaatiolla avainasema. Jos korrelaatio on nolla, ei toisesta muuttujasta ole apua, vaan joudumme tyytymään Y-muuttujan todennäköisimpään arvoon, sen keskiarvoon. Mitä suurempi korrelaatio on, sitä "suurempi oikeus" meillä on poiketa Y:n keskiarvosta X:n osoittamaan suuntaan. Kun korrelaatio kasvaa ykköseksi, ei epävarmuutta enää ole, vaan voimme valita Y:ltä täsmällisen X:ää vastaavan arvon. Kaavan muodossa nämä ajatukset toteutuvat seuraavalla tavalla:

$$Y' = a + bX$$

jossa: Y-pilkku on ennustettu arvo, b on ennustesuoran kulmakerroin, a on vakiotermi eli kohta jossa ennustesuora leikkaa Y-akselin kun X:llä on arvo nolla

Kulmakerroin on määräytyy tulomomenttikertoimen eli korrelaation pohjalta ja vakiotermi taas muuttujien keskiarvojen perusteella.

Yhteydet näihin ovat seuraavat (kulmakerroin ja vakiotermi):

$$b = r\left(\frac{s_y}{s_x}\right) \qquad a = \overline{Y} - b\overline{X}$$

Jos muunnamme arvot ennen tarkastelua Z-pisteiksi yksinkertaistuu ennustekaava siten, että vakiotermi on nolla ja kulmakerroin (myöh. beta) on sama kuin r eli:

$$Z'_Y = r_{xy^*} Z_X$$

Raakapistemäärän ennustesuora ja korrelaatiokerroin sisältävät identtisen informaation.

Meidän esimerkkiaineistossamme olevista muuttujistahan ovat hajonnat, keskiarvot ja korrelaatiot tiedossa, joten voimme helposti tehdä niiden pohjalta arvioita muuttujalta toiselle sekä raakapisteinä että Z-pisteinä. Voimme vaikka ensiksi kysyä, mikä on sellaisen henkilön todennäköinen opintomenestyspistemäärä, jonka matematiikan numero on 8. Äskeiseen kaavaan sijoitettuna tästä tulee raakapisteinä ja Z-pisteinä:

$$Y' = 12.32 + 0.49 * 8 = 16.2$$

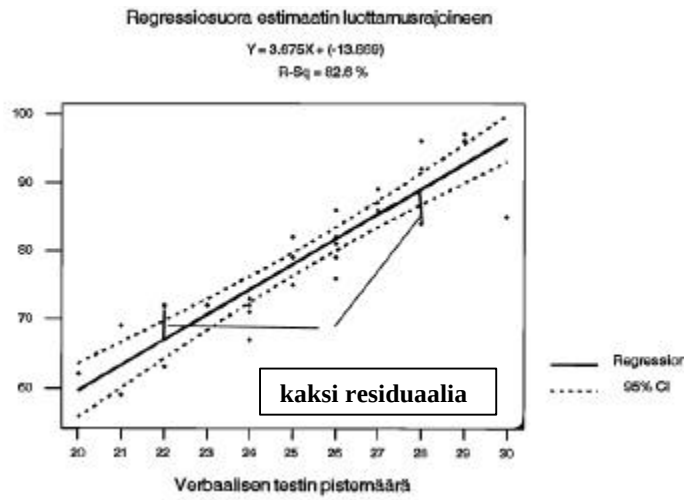
Saatu arvio, 16.20 on siis hiukan opintomenestyksen keskiarvon (15.83) yläpuo1el1a. Tämä on ymmärrettävää, sillä matematiikan numerokin on hiukan oman keskiarvonsa yläpuolella ja muuttujien välinen korrelaatio on positiivinen.

Voisimme seuraavaksi katsoa, mitä tapahtuu, kun korrelaatio on positiivinen ja X-arvo omaa keskiarvoaan pienempi. Kysytään vaikka arviota kielten keskiarvosta, kun verbaalisen testin pistemäärä on 22. Saamme seuraavan laskutoimituksen:

$$Y' = -13.87 + 3.68 * 22 = 66.96$$

Saatu aro 66.96 (oikeastaan 6.969, koska primaarimatriisissa ei ollut desimaalipilkkua) on nyt kielten arvosanan keskiarvon alapuolella.

Pallokuvioissa aikaisemmin näytettiin se seikka, että ennustamisen epävarmuus pienenee korrelaation kasvaessa. Tällaisesta kuviosta näkyy myös se, että ennusteen luottamusrajat eivät ole lineaariset. Poikkeama lineaarisuudesta luottamusrajoissa ei ole kuitenkaan erityisen voimakas, joten sitä ei tarvitse tulosten osalta juuri ottaa huomioon.



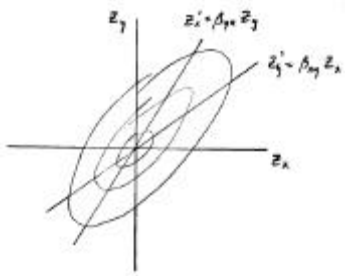
Kuvioon on merkitty kahden pistemäärän osalta myös poikkeama ennustesuorasta eli residuaali. Ennustesuora kulkee pisteparven läpi siten, että tällaisten poikkeamien neliösumma (Sum of Squares) on minimi (pienimmän neliösumman kriteeri).

Yksittäisen havainnon (henkilön) y-pistemäärä jakautuu kahteen osaan: siihen mikä tulee ennusteen (regressiosuoran) kautta ja siihen mikä on poikkeamaa regressiosuorasta joko ylös tai alas eli jäännökseen, residuaaliin. Eta-toiseen -kertoimen tapaan koko pistemääräjoukossa on: Y :n kokonaisvaihtelu SStot, regression selittämä osuus SSreg ja vaille selitystä jäävä satunnaisosa SSres. Korrelaatiokertoimen neliö on suhde SSreg/SStot. Jäännös on osittain satunnaisvarianssia, mittausvirhettä. Osa siitä on systemaattista varianssia, joka ei kuitenkaan käytetyillä muilla (tässä tapauksessa yhdellä) muuttujilla tule selitetyksi. Residuaalipistemäärästä on hyvä muistaa, että sen poikkeamat regressiosuorasta kumoavat toisena. Sen keskiarvo on 0. Residuaaleilla on oma tilastollinen käyttönsä tilanteissa, joissa halutaan tunnettujen tekijöiden osuus poistaa ja jäännöksestä tutkia vieläkö sitä muilla jäljellä olevilla tekijöillä kyetään selittämään. Esim. kovarianssianalyysi on tällainen tekniikka.

Raakapistemääräregressio on varsin harvinainen käytännössä. Yleensähän pistemäärillä on merkitystä vain suhteessa muihin pistemääriin. Onko mitattu arvo keskiarvon ylä- tai alapuolella ja kuinka monta hajonnan mittaa: se on olennaista. Mieleen pitäisi palauttaa Z-pistemäärää koskevat asiat alkeisopinnoista. Asia liittyy myös siihen, että jos jakauma on vähänkään normaalijakauman

suuntainen, niin samalla syntyy myös käsitys pistemäärän suhteellisesta suuruudesta muihin pistemääriin verrattuna. Yhteisjakaumakin voi olla normaali ja sen eri sektoreille sijoittumista voi mieltää visuaalisesti.

Yhden ennustemuuttujan regressiossa z-pisteisiin sovellettu regressiokerroin (beta-kerroin) on yhtä kuin korrelaatiokerroin. Regressioyhtälöstä jää vakiotermi pois. Tällainen regressiosuora kulkee aina origon kautta.



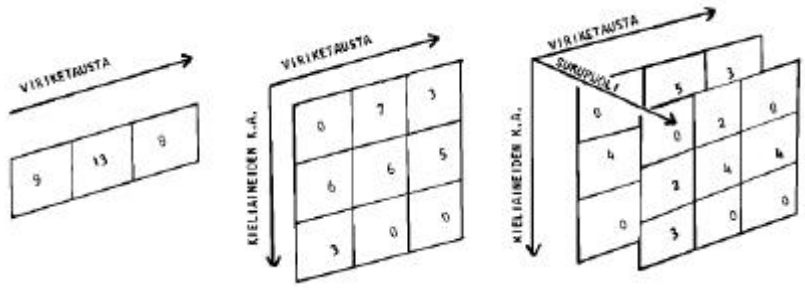
Oheinen käsin hahmoteltu kuvio haluaa vielä muistuttaa siitä, että ennustesuoria on aina kaksi. X:n regressio Y:lle ja Y:n regressio X:lle. b-kertoimet riippuvat hajonnoista. Z-skaalatut kertoimet ovat yhtä suuria keskenään ja samalla korrelaatiokertoimen suuruisia. b-kertoimien geometrinen keskiarvo (ns. keskiverto eli neliöjuuri(b_{xy}*b_{yx}) on suuruudeltaan korrelaatiokerroin. Yleensä asia esitetään yksinkertaistettuna x:n suunnasta y:hyn.

Z-pisteiksi muunnetun muuttujan keskiarvo on tarkastellussa joukossa 0 ja hajonta 1. Koska muuttujien keskiarvo- ja hajontaerot johtuvat triviaaleista asioista (kuten osioiden lukumäärästä, onko etäisyys kouluun kilometreinä vai satoina metreinä ilmaistu, kuukausitulot, viimeisen kuukauden aikana ostettujen kirjojen lukumäärä) niin monimuuttujaisessa tarkastelussa Z- pistemääräinen skaalaus tuo vertailukelpoisuuden erilaisiin kertoimiin. Asia on todella keskeinen.

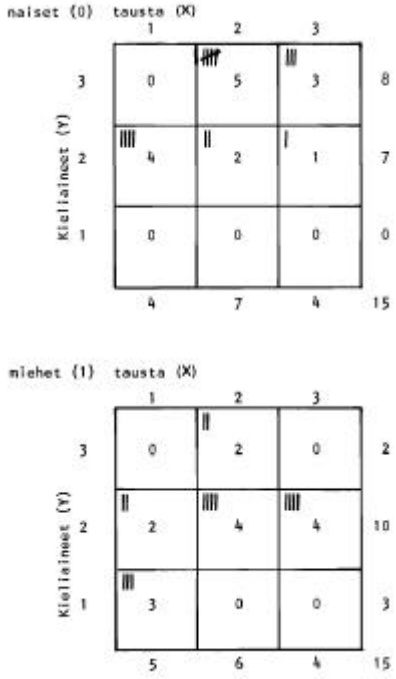
3. Useamman kuin kahden muuttujan yhteyden kuvaus

a) Kolmisuuntaiset ristiintaulukot

Ristiintaulukoinnin ei tarvitse mitenkään välttämättä pysähtyä aiemmin esitettyyn kaksisuuntaiseen taulukkoon; melko usein näkee myös kolmi- suuntaisia ristiintaulukointeja, enempikin on mahdollista. Jotta ajatus kävisi selväksi, on ehkä hyvä miettiä asia alusta alkaen läpi. Ensimmäisenä vaiheena voimme ajatella yksisuuntaista jakaumaa: yhden muuttujan frekvenssit esitetään halutussa määrässä luokkia. Meidän aineistossamme se voisi olla vaikkapa viriketaustan jakauma. Tällöin siis näemme, kuinka monella on huono, keskinkertainen tai hyvä tausta. Kun tuomme tähän yhden suunnan (muuttujan) lisää, saamme kaksisuuntaisen jakauman, ristiintaulukon. Meidän esimerkissämme se voisi olla vaikka viriketaustan ja kieliaineiden keskiarvon yhteyttä kuvaava taulukko, joka on aiemmin esitetty. Tästä siis näemme, miten eri viriketaustat esiintyvät yhdessä eri keskiarvojen kanssa. Meillä on siis tavallaan yksisuuntainen jakauma viriketaustasta kullekin kieliaineiden arvojen luokalle. Samaa logiikkaa seuraten voidaan taulukkoon lisätä jälleen yksi suunta, vaikkapa sukupuoli. Nyt meillä on taustan ja kieliaineiden välinen taulukko molemmille sukupuolille erikseen. Koko prosessi voitaisiin kuvata seuraavasti:



Tarkkaan ottaen on viriketaustan, kieliaineiden keskiarvon ja sukupuolen välinen kolmisuuntainen taulukko seuraavanlainen, "päällekkäiset" taulukot peräkkäin esitettyinä:



Kuvaus tulee tilasto-ohjelmissa näin päin. Vasemmalta ylhäältä luokkien arvot alkavat kasvaa molemmissa muuttujissa kohti oikeaa alakulmaa.

```
Muuttuja: sukupuoli = 0 (naiset)

Kie-
liaineiden kes-
kiarvo (Y)  Viriketausta (X)
            1          2          3        All

 1          0          0          0          0
 2          4          2          0          6
 3          0          5          4          9

 All        4          7          4         15

Muuttuja: sukupuoli = 1 (miehet)

            1          2          3        All

 1          3          0          0          3
 2          2          4          4         10
 3          0          2          0          2

 All        5          6          4         15
```

Tällaisten ristiintaulukoiden avulla voidaan tulosta täsmentää: missä olosuhteissa yhteys erityisesti esiintyy ja on voimakas. Kolmannen muuttujan mukaan ottaminen voi myös hävittää riippuvuuden. Kahden muuttujan välistä yhteyttä

elaboroidaan (täsmennetään missä olosuhteissa riippuvuus esiintyy eli spesifioidaan ja tulkitaan) ottamalla mukaan kolmas muuttuja (joskus hiukan erikoisesti testimuuttujaksi nimitetty muuttuja, usein ns. taustamuuttuja).

Näin tarkastellen saamme yhteyksiin taas hieman lisää valaistusta. Voimme todeta, että sukupuolittaiset viriketaustan jakaumat (marginaali- eli reunafrekvenssit taulukoiden alla) ovat lähes samat. Kieliaineiden arvosanat taas ovat naisilla yleensä paremmat (reunajakaumat oikealla). Miehiä ja naisia on aineistossa yhtä paljon (numerukset oikeassa alanurkassa). Kaikki aineistossa olevat huonoimman kieliarvosanan saaneet ovat miehiä, kun taas vain kaksi parhaista on miehiä. Jostakin syystä näyttää yhteys olevan miesten joukossa käyräviivaisempi kuin naisten; vaikka keskimääräisen viriketaustan sarakkeessa on kaksi parhaan arvosanan saanutta, ei heitä enää ole oikeanpuoleisessa sarakkeessa.

Useampiulotteisessa taulukoinnissa on muistettava, että aineistojen on oltava suhteellisen suuria. Mitä useampiin erilaisiin ryhmiin sama aineisto jaetaan, sitä vähemmän tapauksia enää riittää kuhunkin ruutuun. Taulukoista tulee helposti liian "laihoja", jolloin niiden merkitys ja uskottavuus kuvauksena vähenee.

Loglineaarinen mallinnus tarjoaa taloudellisemman keinon ja ehkä myös teorialäheisemmän tavan useiden dikotomisten tai trikotomisten muuttujien yhteyksien tarkastelemiseen ja yhteyksiä koskevien hypoteesien testaamiseen.

b) Osittaiskorrelaatio

Tutkimuksessa yleensä, mutta varsinkin kokeellisessa tutkimuksessa, pyritään saamaan tutkittavaan asiaan kuulumattomien tekijöiden vaikutus mahdollisimman pieneksi, ts. ne pyritään kontrolloimaan tai vakioimaan. Muutenhan olisi tavattoman vaikea tietää, mistä saadut yhteydet itse asiassa johtuvat. Oletetaan vaikka, että olemme kiinnostuneita verbaalisen (kielellisen) ja spatiaalisen lahjakkuuden (avaruustajun) välisestä yhteydestä. Käytämme koehenkilöinä vaikkapa peruskoulun viidesluokkalaisia ja teemme siis heille kumpaakin lahjakkuuden faktoria mittaavat testit. Tulos voisi olla vaikkapa .40 korrelaatio testien välillä. Tämähän :merkitsee kohtalaisen selvää yhteyttä: keskimäärin me-

nestyy toisella testillä samantapaisesti kuin toisellakin, siis samat henkilöt pyrkivät olemaan hyviä molemmissa tai huonoja molemmissa, vaikkakin poikkeuksia on.

Voimmeko nyt sitten olla varmoja siitä, että verbaalisen ja spatiaalisen kyvyn välillä on tällainen yhteys? Tarkkaan ottaen emme. Jos testit tehdään riittävän suurelle koehenkilöjoukolle, voimme kyllä luottaa siihen, että testisuoritusten välillä on saadun kaltainen yhteys. Jotta voisimme sanoa ko. kykyjen välillä vallitsevan tällaisen yhteyden, pitäisi verbaalisen testin mitata vain verbaalista kykyä ja spatiaalisen vain spatiaalista, mikä tuskin koskaan on mahdollista. Lopullinen suoritus testissä koostuu useista komponenteista, joiden osuudet kokonaisuudesta vaihtelevat.

Koska koehenkilöt ovat jokseenkin samanikäisiä, on tässä tapauksessa eniten epäilyksiä herättävä häiritsevä seikka yleinen älykkyys. Luultavasti sekä verbaalinen että spatiaalinen testi sisältävät kumpikin omalta osaltaan myös yleisen älykkyyden vaikutusta. Tällainen yhteinen komponentti aiheuttaa positiivista korrelaatiota muuttujien välille. Saattaa olla, että saamamme korrelaatio aiheutuukin yleisen älykkyyden osuudesta eikä ole osoitus verbaalisen ja spatiaalisen lahjakkuuden yhteydestä sellaisenaan. Tämän seikan selvittämiseksi tulisi yleisen älykkyyden osuus poistaa mittaustuloksista, se tulisi vakioida.

Vakiointi voidaan tehdä kahdella periaatteessa erilaisella tavalla. Suoraviivaisin ja ymmärrettävin, mutta työteliäs tapa on yksinkertaisesti hankkia yleiseltä älykkyydeltään samanlaisia koehenkilöitä. Toinen on halutun muuttujan vakioiminen tilastollisin keinoin aineistossa, jossa se ei alunperin ole vakio. Tällainen menetelmä on osittaiskorrelaation laskeminen. Osittaiskorrelaation kaava on seuraava:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r^2_{13})\ (1-r^2_{23})}}$$

$r_{12.3}$ tarkoittaa muuttujien 1 ja 2 välistä korrelaatiota, kun muuttuja 3 on vakioitu, r_{12} on muuttujien 1 ja 2 korrelaatio jne. Osittaiskorrelaation laskemiseksi tarvitaan siis kaikkien muuttujien väliset korrelaatiot.

Omassa esimerkkiaineistossamme voimme todeta, että naiset ovat menestyneet opinnoissaan miehiä paremmin (sukupuolen ja opintomenestyksen korrelaatio on -.39). Verbaalisen lahjakkuuden ja opintomenestyksen korrelaatio on huomattavan korkea, .81, kun verbaalinen lahjakkuus on mitattu testillä. Voidaan kysyä, johtuuko naisten menestys opinnoissaan juuri siitä, että he ovat kielellisesti lahjakkaita, vai onko jokin muu tekijä merkittävästi mukana vaikuttamassa. Tähänhän voidaan saada vastaus laskemalla sukupuolen ja opintomenestyksen korrelaatio pitämällä kielellinen lahjakkuus vakiona, ts. tutkitaan, mikä olisi sukupuolten opintomenestys, jos heidän verbaalinen lahjakkuutensa olisi sama. Korrelaatiomatriisista voimme poimia tarpeelliset korrelaatiot. On helpointa merkitä muuttujia symboleilla 1, 2 ja 3 kun kolmas on vakioitava muuttuja. Tällöin symbolit ovat samat kuin kaavassa, eikä sekaannuksia helposti synny. Saamme seuraavat korrelaatiot:

	1	2
1 sukupuoli		
2 opintomenestys	—.39	
3 verbaalinen lahjakkuus	—.45	.81

Laskutoimituksesta tulee seuraavanlainen:

$$r_{12.3} = \frac{-.39 - (-.45 \cdot .81)}{\sqrt{[1-(-.45^2)]\ [1-.81^2]}}$$

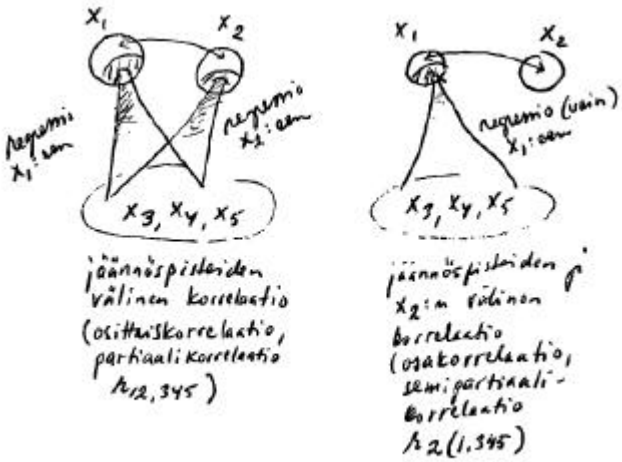
$$= \frac{-.02}{.52} = -.04$$

Siis alkuperäinen sukupuolen ja opintomenestyksen välinen korrelaatio, -.39, pienenee -.04:ään, kun verbaalinen lahjakkuus vakioidaan. Näyttää siis siltä, että sukupuolten välinen ero opintomenestyksessä johtuu juuri verbaalisesta lahjakkuudesta; jos se vakioidaan, häviää yhteys lähes kokonaan.

Osittaiskorrelaatioita voidaan laskea myös pitämällä useita muuttujia vakioina. Näin kuitenkin harvoin varsinkaan käsin laskiessa tehdään, mutta esitettäköön kuitenkin malliksi muuttujien 1 ja 2 osittaiskorrelaation kaava, kun muuttujat 3 ja 4 on pidetty vakioina:

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{(1-r^2_{14.3})\ (1-r^2_{24.3})}}$$

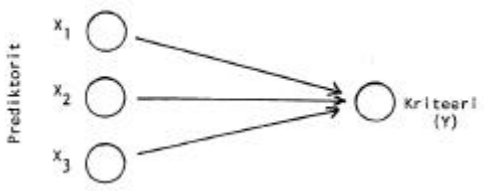
Tämä on ns. toisen asteen osittaiskorrelaatio, koska sen lähtöarvoiksi tarvitaan edellä esitettyjä ensimmäisen asteen osittaiskorrelaatioita.



Edelliset kuviot havainnollistavat osittaiskorrelaation j a osakorrelaation käsitteitä. Osa informaatiosta voidaan poistaa (vakioida) joko molemmista tai toisesta muuttujasta. Osakorrelaation käsite osoittautuu erittäin tärkeäksi asiaksi regressioanalyysin yhteydessä, kun selittäviä (ennustavia) muuttujia
on kaksi tai enemmän. Palaapa tähän kuvioon regressioanalyysin jälkeen. Kun muuttujaa 1 selitetään muuttujilla 3,4 ja 5, pystytään sen vaihtelusta (variansista) tilastollisesti selittämään tietty osuus (=yhteiskorrelaation eli multippelikorrelaation neliö). Jos selitettävien muuttujien joukkoa täydennetään vielä muuttujalla 2, se tuo lisää selitystä osakorrelaation eli semipartiaalikorrelaation neliön verran.

c) Regressioanalyysi

Kuvitellaanpa nyt, että olemme päässeet (tai joutuneet!) jonkin oppilaitoksen johtoon ja saaneet tehtäväksemme suunnitella pyrkijöiden valintojen kehittämistä. Meillä on mahdollisuus pyytää pyrkijöistä tarpeellisia taustatietoja sekä tehdä heille joitakin testejä. On kuitenkin vaikea tietää, mitkä tiedot ovat käyttökelpoisimpia. Jotkut voivat olla parempia valinnassa kuin toiset, joillakin voi olla niin suurta päällekkäisyyttä, että osa on turhaa jne. Koska haluamme sellaisia henkilöitä, jotka tulevat menestymään opinnoissaan, on tehtävänä itse asiassa ennustaa opintomenestystä (kriteeriä) käytettävissä olevien tietojen (prediktorien) avulla:



Kuten aiemmin esitetyssä tapauksessa, jossa haetaan parasta ennustetta yhden tunnetun muuttujan avulla, täytyy tässäkin olla käytettävissä koehenkilöjoukko, jolla on hankittu mitat kaikilla muuttujilla. Tässä esimerkissä se siis merkitsee pyrkijöiden joukkoa, jolle on tehty testit, jolta on kerätty taustatiedot, ja joka on ehtinyt opiskella niin, että opintomenestys on voitu kohtuullisella luotettavuudella arvioida. Näitä tietoja käytetään apuna ennusteen teossa uusille pyrkijöille. Selektion ongelma on tärkeä havaita jo nyt.

Ennusteen voi tietysti tehdä monella tapaa, vaikkapa vain laskemalla kaikkien mittojen (muuttujien) summan, mutta tässä tapauksessa me haemme nimenomaan parasta ennustetta, sellaista painotettua yhdistelmää, joka selittää kriteeristä mahdollisimman paljon. Prediktorien erilaisuuden (hajonnat) ja päällekkäisyyden (korrelaatiot) takia muuttujat täytyy ottaa huomioon eri määrin, jollekin prediktorille annetaan ennusteessa suurempi paino kuin toiselle. Kussakin tapauksessa on löydettävissä tietyt painokertoimet, jotka aikaansaavat parhaan mahdollisen ennusteen. Näiden painokertoimien löytämiseen, samoin kuin tehdyn ennusteen hyvyyden arviointiin, soveltuu (multippeli) regressioanalyysi. Sanalla "multippeli" viitataan siihen, että prediktoreita on useita, mutta koska näin kaikissa mielekkäissä sovellutuksissa on, voimme puhua pelkästä regres-

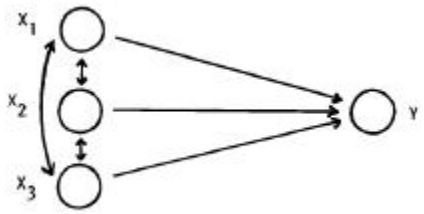
sioanalyysistä. Voimme siis tässä alustavasti määritellä regressioanalyysin tilastolliseksi menetelmäksi, jolla haetaan parasta mahdollista selittävien muuttujien (prediktorien) painotettua yhdistelmää ennustettaessa yhtä selitettävää muuttujaa (kriteeri).

Korrelaation yhteydessä on jo todettu, että muuttujan selitysosuus toisesta voidaan ilmoittaa selitysosuutena korottamalla korrelaatiokerroin toiseen. Esimerkiksi omassa esimerkkiaineistossamme selittäisi sukupuoli täten opintomenestystä -.392 = 15.2 %. Tässähän näyttäisi olevan keino selitysosuuksien hankkimiseksi. Voimme katsoa, mitä tapahtuu, kun tällä tavoin selitämme esimerkkiaineistomme opintomenestystä muilla siinä olevilla muuttujilla. Oheen on kerättyopintomenestyksen ja muiden muuttujien väliset korrelaatiot sekä niiden neliöt:

	r	selitys %
sukupuoli	—.39	15.2
viriketausta	.76	57.8
verbaalinen testi	.81	65.6
järkeilytesti	.32	10.2
kielten k.a.	.83	68.9
matematiikan nro	.29	8.4
	yhteensä	226.1%

Olemmeko nyt onnistuneet erikoisen hyvin, kun saimme "selitetyksi" peräti 226.1 prosenttia opintomenestyksen vaihtelusta? Varmasti emme, onhan täysin epäloogista sanoa, että jostakin ilmiöstä selitetään paljon enemmän kuin se kokonaisuudessaan. Missä sitten on vika, eikö korrelaation neliö olekaan selitysosuuden mitta? Tähän voisi vastata monellakin tapaa, mutta toteamme tässä, että se on aivan kuten aiemmin on esitettykin, mutta tämä pätee vain yhden selittäjän tapauksessa. Kun selittäjiä on useita, on niiden välillä tavallisesti korrelaatiota, ts. selittäjät ovat osittain päällekkäisiä. Kun yhden selittäjän osuus on määritetty, on itse asiassa käytetty jo pala toisestakin, eikä tätä osuutta saa enää käyttää uudelleen. Jos esimerkiksi meidän aineistossamme selitämme opintomenestystä kielten keskiarvolla, ei verbaalinen testi enää paljoa lisää ennusteen hyvyyttä, koska kielten keskiarvo ja verbaalinen testi ovat suureksi osaksi mittoja samasta asiasta. Mitä suurempi siis on prediktorien, selittäjien, välinen korrelaatio, sitä suurempaa on niiden päällekkäisyys ja sitä vähemmän auttaa enää uusien prediktorien käyttö. Meidän tekemämme virhe oli siis prediktorien välisten korrelaatioiden huomiotta jättäminen. Edellisessä kuviossa pitäisi olla kaksipäiset

nuolet X-muuttujien välillä kuvaamassa sitä, että ennustemuuttujien väliset korrelaatiot on otettu huomioon.



Korrelaatiot otetaan huomioon painotuksen kautta siten, että selitys on maksimaalinen. Myöhemmin havaitset, että selitykseen käytetään käsitettä Y' (estimaatti), joka on selitettävien muuttujien painotettu summa.

Kun haluamme tietää todellisen selitetyn osuuden, jossa prediktorien päällekkäisyys on otettu huomioon, on laskettava multippelikorrelaation neliö, R2. Kuten saattaa jo arvatakin, siihen tarvitaan lähtötiedoiksi prediktorien ja kriteerin väliset sekä prediktorien väliset korrelaatiot. Kahden prediktorin tapauksessa kaava on seuraava:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

Yleinen kaava on:

$$R^2 = \beta_{12.3} * r_{12} + \beta_{13.2} * r_{13}$$

$$\text{eli yleisemmin } R^2 = \sum \beta r$$

Kriteeriä on merkitty ykkösellä ja prediktoreita kakkosella ja kolmosella. Voimme nyt soveltaa tätä äsken esitettyyn kysymykseen, miten kielten keskiarvo ja verbaalinen testi yhdessä selittävät opintomenestystä. Korrelaatiot ovat seuraavat:

	1	2
1 opintomenestys		
2 verbaalinen testi	.81	
3 kielten k.a.	.83	.91

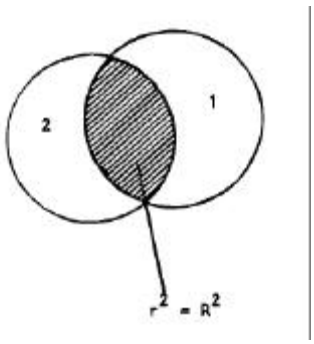
Kaavaan sijoittamalla ja laskemalla tulemme seuraavaan tulokseen:

$$R^2_{1.23} = \frac{.81^2 + .83^2 - 2 \cdot .81 \cdot .83 \cdot .91}{1 - .91^2}$$

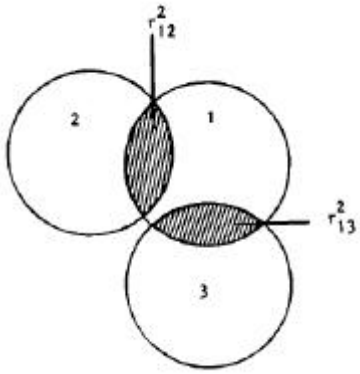
$$= \frac{1.35 - 1.22}{1 - .83} = .76$$

Kielten keskiarvo ja verbaalinen testi siis selittävät opintomenestystä yhteensä 76 %. Kun kielten keskiarvo selittää jo $.832^2$ = 69 %, ei siis verbaalisen testin lisääminen selitykseen enää lisää kokonaisselitysosuutta kuin 7 %. Tämähän johtui prediktorien, verbaalisen testin ja kielten keskiarvon, välisestä korkeasta korrelaatiosta (.91).

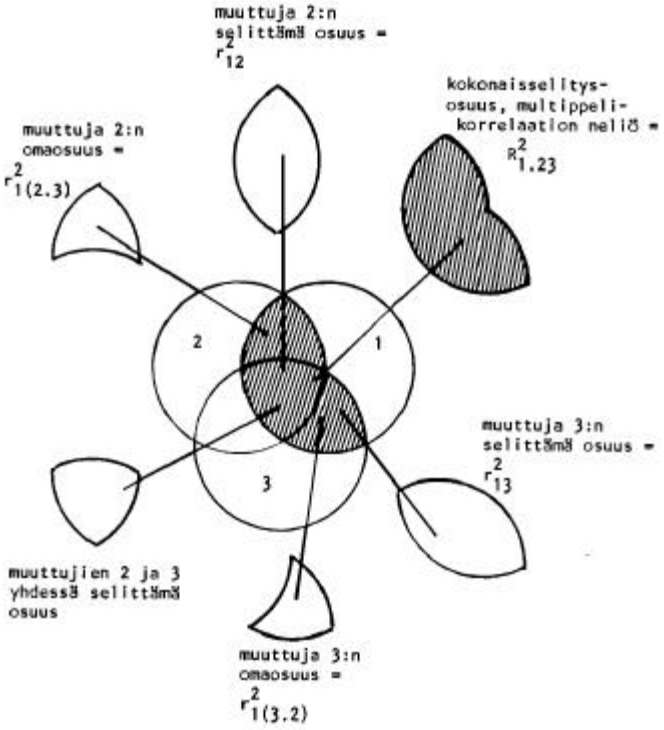
Regressioanalyysiin liittyvä terminologia selviää ehkä parhaiten kuvallisesta esityksestä, jossa ympyrät kuvaavat kunkin muuttujan vaihtelua. Päällekkäin olevat osat ovat tällöin yhteistä vaihtelua, sitä, jonka muuttuja selittää toisesta. Merkitsemme äskeisen esimerkin mukaan kriteeriä ykkösellä ja selittäjiä, prediktoreita, kakkosella ja kolmosella. Yhden selittäjän tapaus on selkeä ja yksinkertainen. Kuten useasti on jo todettu, selitettyosuus on muuttujien välisen korrelaation neliö, r^2. Koska yhden prediktorin tapauksessa tämä on myös kaikki, mitä on selitetty, on se samalla multippelikorrelaation neliö, $r^2 = R^2$. Samoin sama osuus on muuttuja kakkosen yksin selittämä osuus, omaosuus (eli semipartiaalikorrelaation neliö, osakorrelaation neliö):



Lähes yhtä selkeä on kahden (tai useamman) korreloimattoman prediktorin tapaus. Kumpikin prediktori selittää kriteeristä oman osuutensa, jotka samalla ovat näiden muuttujien omaosuuksia. Koko selitetty vaihtelu, multippelikorrelaation neliö, on erillisten selitysosuuksien summa. Tällöin, ja vain tällöin, pätee se laskutapa, jonka alussa esitimme, jossa laskettiin suoraan korrelaatioiden neliöiden summa kokonaisselitykseksi.



Tärkein tilanne, juuri se mihin regressioanalyysiä varsinaisesti tarvitaan, on kahden (tai useamman) korreloivan prediktorin tapaus. Nyt saavat korrelaatioiden neliöt, multippelikorrelaation neliö ja omaosuudet kaikki oman, toisistaan poikkeavan merkityksensä, joita ei saa sekoittaa keskenään. Jotta jokainen osuus varmasti täsmällisesti selviäisi, on ne piirretty kuviosta "ulos" silläkin uhalla, että kuva ensin näyttää monimutkaiselta:



Kuvan tutkiminen selvittää monta oleellista asiaa. Kokonaisselitysosuus ei nyt ole erillisten osuuksien summa. Jos lasketaan yhteen prediktori en ja kriteerien välisten korrelaatioiden neliöt saadaan liian suuri osuus, koska prediktorien yhdessä selittämä osuus tulee mukaan kaksi kertaa. Jos taas lasketaan yhteen kummankin selittäjän omaosuudet, se mitä ne yksin selittävät, tulee summasta liian pieni, koska yhteinen osuus ei ole mukana lainkaan.

Multippelikorrelaation neliö on siis omaosuuksien ja yhteisen osuuden summa. Prediktorin omaosuus on se osuus kokonaisselityksestä, joka tulee mukaan lisättäessä prediktori analyysiin. Toisin sanoen, prediktorin lisääminen kasvattaa multippelikorrelaation neliötä tämän prediktorin omaosuuden verran. Kun aiemmin selitimme opintomenestystä kielten keskiarvolla, saimme selitysosuudeksi 69 %. Verbaalisen testin lisääminen selitysmalliin nosti kokonaisselityksen 76 %:iin. Näiden ero, 7 %, on verbaalisen testin omaosuus.

Selitysosuuksien lisäksi tuottaa regressioanalyysi myös painokertoimet, joilla kukin yksityinen pistemäärä on kerrottava, jotta kokonaisselitys olisi mahdollisimman hyvä, ts. jotta saataisiin mahdollisimman suuri multippelikorrelaatio. Jos käytetään raakapisteitä sellaisinaan, kuten esim. alussa esitetyssä havaintomatriisissa on, ovat painokertoimet ns. b- kertoimia (osittaisregressiokertoimia). Jos taas pisteet on ensin standardoitu, ts. on laskettu Z-pisteet, joiden keskiarvo on nolla ja hajonta yksi, ovat kertoimet beta-kertoimia (standardoituja osittaisregressiokertoimia). Beta-kertoimet saadaan muuttujien välisistä korrelaatioista seuraavalla tavalla:

$$\beta_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}$$

$$\beta_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}$$

Muuttuja ykkösen ollessa kriteeri, on $beta_{12.3}$ kakkosmuuttujan ja $beta_{13.2}$ kolmosmuuttujan painokerroin. Kertoimet ovat vertailukelpoisia, koska muuttujat on standardoitu. Jos jatkamme edelleen esimerkkiä, jossa opintomenestystä selitettiin verbaalisella testillä ja kielten keskiarvolla, saamme seuraavat laskut:

$$\beta_{12.3} = \frac{.81 - .83 \cdot .91}{1 - .91^2} = \frac{0.54}{.172} = .31$$

$$\beta_{13.2} = \frac{.83 - .81 \cdot .91}{1 - .91^2} = \frac{.093}{.172} = .54$$

Muuttuja kakkosen (verbaalinen testi) beta-kerroin on siis .31 ja muuttuja kolmosen (kielten keskiarvo) beta-kerroin on .54. Tämä tarkoittaa sitä, että jos meillä on kunkin henkilön verbaalisen testin ja kielten keskiarvon pistemäärä Z-pisteinä, saamme hänelle parhaan ennus- teen opintomenestyksessä Z-pisteinä kertomalla pisteet beta-kertoimilla ja laskemalla ne yhteen. Kaavan muodossa tämä on:

$$Z'_1 = \beta_{12.3} Z_2 + \beta_{13.2} Z_3$$

Voimme ottaa esimerkiksi havaintomatriisimme ensimmäisen henkilön. Hänen verbaalisen testin pistemääränsä oli 22 ja kieliaineiden keskiarvonsa 63 (ilman desimaalipilkkua). Jotta beta-kertoimia voisi soveltaa, on nämä ensin muutettava Z-pisteiksi vähentämällä niistä ko. muuttujan keskiarvo ja jakamalla erotus standardipoikkeamalla:

$$Z_2 \text{ (verb. testi)} = \frac{22 - 25.47}{2.57} = -1.35$$

$$Z_3 \text{ (kieliaineet)} = \frac{63 - 79.73}{10.39} = -1.61$$

Tällä henkilöllä tuntuu menevän heikonlaisesti: molemmat arvot ovat reilusti yli yhden keskihajonnan verran keskiarvon alapuolella. Lienee siis odotettavissa, että myös opintomenestyksen standardiarvo olisi negatiivinen. Jos nyt siis oletamme, että hänellä ei olisi opintomenestyksen mittaa tai haluaisimme tutkia, onko hänen menestymisensä ennusteen mukaista, voimme laskea odotetun opintomenestyksen standardiarvon:

$$Z_1' = .31\,(-1.35) + .54\,(-1.61) = -1.29$$

Ennuste on odotusten mukaisesti selvästi keskiarvon alapuolella. Tässä tapauksessa menestys on kuitenkin ollut vielä ennustettakin huonompaa, koska ko. henkilön opintomenestys on standardipisteinä -1.54.

Beta-kertoimet ovat siitä käteviä, että ne ovat suoraan toisiinsa verrattavissa. Esimerkiksi tässä esimerkissä näemme, että kieliaineiden keskiarvo on hieman parempi ennustaja kuin verbaalinen testi. Monissa käytännön tilanteissa, esim. oppilasvalinnoissa, emme kuitenkaan ole vain kiinnostuneita prediktorien tehokkuudesta, vaan tarvitsemme tietoa siitä, millä arvoilla primääripisteitä sellaisinaan on painotettava parhaan ennusteen saamiseksi. Nämähän olivat b-kertoimia ja ne saadaan beta-kertoimista yksinkertaisesti painottamalla niitä muuttujien hajontojen suhteella:

$$b_{12.3} = \frac{S_1}{S_2} * \beta_{12.3} = \frac{1.84}{2.57} * .31 = .22$$

$$b_{13.2} = \frac{S_1}{S_3} * \beta_{13.2} = \frac{1.84}{10.39} * .54 = .096$$

Näillä luvuilla siis kerrotaan kunkin henkilön pistemäärät, jolloin saadaan arvio opintomenestyksen pistemäärästä. Aivan vielä emme kuitenkaan ole valmiita tätä tekemään. Koska kaikki pisteet ovat nyt standardoimattomia, voivat niiden suuruusluokat olla mitä tahansa. Jotta vastaus saataisiin sillä asteikolla, jolla kriteeri on mitattu, siis jotta lukujen suuruusluokka olisi oikea, tarvitaan vielä vakio. Tämä vakio, jota merkitään a:lla, lisätään b-kertoimilla kerrottuihin pistemääriin, jolloin regressio- yhtälö saa seuraavan muodon:

$$X_1' = a + b_{12.3} * X_2 + b_{13.2} * X_3$$

Vakion laskemiseen tarvitaan prediktorien ja kriteerin keskiarvot sekä b-kertoimet:

$$a = \overline{X}_1 - b_{12.3} * \overline{X}_2 - b_{13.2} * \overline{X}_3 = 2.57$$

Nyt meillä on kaikki tarvittavat tiedot ja voimme laskea vaikkapa havaintomatriisimme kolmelle ensimmäiselle henkilölle odotetun opintomenestyspistemäärän (13.46. 16.16 ja 18.16). Laadi yhtälö, jolla ne saadaan.

Kun vertaamme näitä (Y') taulukon todellisiin (Y) arvoihin, voimme todeta niiden menneen kohtalaisen hyvin kohdalleen. Suurin ero on kolmannella koehenkilöllä, jonka suuri kieliaineiden keskiarvo teki ennusteesta hieman liian suuren. On myös muistettava, että nämä ennusteet perustuivat vain kahden prediktorin käyttöön. Useamman prediktorin regressioanalyysi on kuitenkin laskennallisesti raskas ja tehdään yleensä poikkeuksetta valmiilla tilasto-ohjelmilla. Edellä esitetyt laskutoimenpiteet on kuitenkin aiheellista käydä läpi, ettei koneella tuotettu tulostus muodostuisi mystiseksi tempuksi, jonka joutuu ottamaan kritiikittömästi sellaisenaan.

Voimme nyt myös esittää primääriaineistostamme tehdyn regressioanalyysin tulostuksen, kun opintomenestystä on selitetty kaikilla muilla muuttujilla. Ohjelmat tulostavat yleensä tärkeimmät arvot seuraavasti:

$R^2 = .81$ Vakio (a) = 1.64

muuttuja	b-kerroin	beta-kerroin	omaosuus
5 (kieliaineiden k.a.)	.11	.64	.06
2 (viriketausta)	.78	.32	.03
1 (sukupuoli)	—.76	—.21	.02
4 (järkeilytesti)	.15	.15	.01
6 (matem. nro)	.23	.14	.01
3 (verb. testi)	—.12	—.16	.00

Yhteensä saatiin siis opintomenestyksestä selitetyksi 81 %. Paras selittäjä oli kieliaineiden keskiarvo, mutta prediktorien voimakkaan päällekkäisyyden vuoksi oli senkin omaosuus vain 6 %. Tämän päällekkäisyyden takia on myös osa näistä prediktoreista turhaa, totesimmehan esimerkiksi aiemmin, että pelkästään verbaalisella testillä ja kieliaineiden keskiarvolla saadaan selitetyksi jo 76 % opintomenestyksestä. Paras kahden muuttujan kombinaatio on kieliainei-

den keskiarvo ja matematiikan numero, joka selittää 77.4 %. Itse asiassa analyysin olisi voinut lopettaa tähän, koska loppujen prediktoreiden mukaantulo lisää selitystä niin vähän, että se voi melkein yhtä hyvin olla pelkkää sattumaa. Muuttujan lisääminen ei koskaan voi pienentää jo saavutettua selityksen astetta.

Ohjelmissa voi yleensä valita pakollisen tai valikoivan "mallin " välillä. Pakollisessa mallissa ilmoitetaan ne prediktorit, jotka halutaan mukaan, ja kone laskee tuloksen koko tälle joukolle. Valikoiva malli ottaa mukaan prediktorin kerrallaan, aina järjestyksessä sen, joka edelliseen tilanteeseen verrattuna kasvattaa selitysastetta eniten, kunnes halutut muuttujat ovat kaikki mukana. Valikoiva malli on tutkijalle "helppo", koska ei ole tarpeen edeltä käsin arvioida, minkälaisia tulokset olisivat. Samalla se on vaarallinen, koska se aiheuttaa helposti "kai sieltä jotakin tulee" -tyyppistä tutkimusta, jossa mikä tahansa aineisto syötetään ohjelmaan toivoen, että jotakin mielenkiintoista ilmaantuisi. Pakollinen malli on siis usein tutkimusmielessä "terveempi"; tutkija testaa oletetun selityksen todellisuutta eikä ole "ohjelman armoilla".

Mallista riippumatta pyrkivät regressioanalyysin tulokset olemaan "liian hyviä", yliestimaatteja todellisesta. Otokselle räätälöidyt painokertoimet tuottavat maksimin vain otoksessa, jossa ne on laadittu. Jos hankimme painokertoimet yhden otoksen perusteella ja käytämme sitten niitä toisessa aineistossa, jää selitettyosuus melko varmasti pienemmäksi kuin alkuperäisellä materiaalilla. Tämä johtuu siitä, että analyysi pyrkii koko ajan maksimoimaan selityksen, "ottamaan otoksesta irti kaiken mahdollisen". Tähän sisältyy myös virhettä ja vain ko. otokselle tyypillistä vaihtelua, joka toisessa otoksessa tuskin toistuu juuri sellaisena. Jos me siis hankkisimme toisen otoksen ja tekisimme opintomenestyksen ennusteet äsken esitetyillä kertoimilla, multippelikorrelaation neliö jäisi melko varmasti alle 81 %:n. Tällaista yhdellä aineistolla saatujen tulosten tarkistamista toisen aineiston avulla nimitetään ristiinvalidoinniksi. Sen tekeminen silloin kun se on mahdollista, on erittäin suositeltavaa, koska se lisää tulosten uskottavuutta huomattavasti.

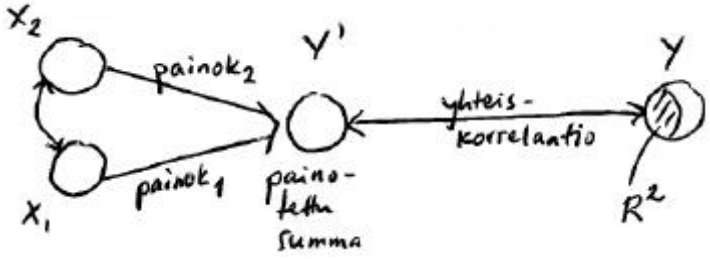
Kuten kaikki monimuuttujamenetelmät, käy regressioanalyysikin nopeasti epäluotettavaksi, jos tapausten määrä (joka yleensä merkitsee siis tutkittujen henkilöiden määrää), on liian pieni verrattuna muuttujien määrään. Numeruksen pitäisi olla huomattavasti, mieluummin monta kertaa suurempi kuin muuttujien lukumäärä. Ääritapaus on tilanne, jossa muuttujia ja tutkittuja tapauksia on yhtä paljon. Sattumanvarainen, vain tälle otokselle tyypillinen vaihtelu kasvaa tällöin

sellaisiin mittoihin, että kokonaisselitysosuus on tällaisessa tilanteessa aina 100 %. Ohjelmat tulostavat myös vapausasteiden lukumäärään perustuvan, korjatun arvion. Sekään ei ota huomioon sitä, että selittävät muuttujat on mahdollisesti valittu laajasta muuttujajoukosta. Käyttäjä joutuu itse huolehtimaan omien ratkaisujensa pitävyydestä.

Tässä vaiheessa voidaan todeta, että ei regressioanalyysi liitykään kovin paljon lopulta ennustamiseen. Ennustettua pistemäärää on käytetty vain välineenä, jotta saataisiin selville miten kohdemuuttujan varianssi muodostuu. Tällaiseen tilastolliseen kuvaamiseen ja selittämiseen regressioanalyysi yleensä tutkimuksissa jää. Aito ennustaminen olisi sitä, että saatua regressioyhtälöä sovellettaisiin tapauksiin, josta ei vielä tiedettäisi heidän todellisia Y-pistemääriään.

Käsiteltävän käytännön ongelman kannalta olisi siis vielä matkaa siihen, miten valinnat oppilaitokseen olisi suoritettava.

Muutama hyödyllinen seikka vielä. Jos kuvataan koko regressioketjua, niin voidaan mieltää, että Y' :n ja Y :n välinen korrelaatio on yhteiskorrelaatio eli multippelikorrelaatio. Regressioyhtälö maksimoi sen (ja samalla minimoi jäännösvaihtelun neliösumman, pienimmän neliösumman kriteeri):



Yksittäisen selittävän muuttujan kohdalla sen beta-kerroin ja suora korrelaatio kriteeriin yleensä ovat etumerkiltään samat, mutta niin ei välttämättä ole. On myös mahdollista, että beta on kohtuullinen vaikka suora korrelaatio selityksen kohteeseen on pieni. Kokonaisuus muodostuu selkeäksi silloin, kun selittävien muuttujien välillä ei ole korkeita korrelaatioita. Korkeat korrelaatiot selittävien muuttujien kesken (eli ns. multikollineaarisuus) tekee asian usein vaikeasti hal-

littavaksi. On tietenkin mahdollista esim. sellainen tilanne, että kaksi ennustemuuttujaa korreloi positiivisesti keskenään, mutta toinen negatiivisesti kriteeriin ja toinen taas positiivisesti. Korrelaatiomatriisin korrelaatioita täytyy siis myös tarkastella alkuarvoina eikä ottaa vain regressioanalyysin lopputulosta sellaisenaan tutkimustulokseksi.

Huomaat tulevassa faktorianalyysi-luvussa, että on hyödyllistä rinnastaa faktorit korreloimattomiksi ennustemuuttujiksi ja kommunaliteetti yhteiskorrelaation neliöksi. Samoin faktorilatauksen voi suoraan rinnastaa muuttujan ja faktorin väliseen korrelaatioon. Kun nämä kaksi monimuuttuja-analyysia ymmärtää yhdessä ja erikseen, on käsitejärjestelmä tullut hyvin opittua.

d) Faktorianalyysi

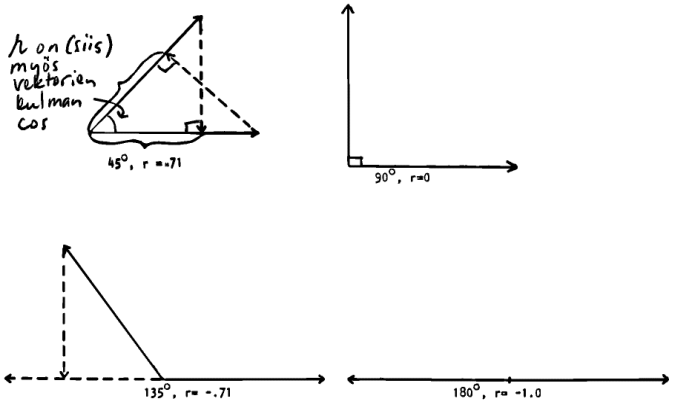
Suomalaista psykologista ja kasvatustieteellistä tutkimuksen tekoa on leimannut runsas faktorianalyysin käyttö. Traditio syntyi Yrjö Ahmavaaran ja Touko Markkasen työn myötä jo 1950-luvu1la. Toivo Vahervuo ja sittemmin vaikkapa Veikko Heinonen veivät perinnettä eteen päin. Sosiologiassa on havaittavissa samankaltaisia piirteitä. Amerikkalainen vaikutus oli voimakasta käyttäytymistieteiden metodiikassa toisen maailmansodan jälkeen.

Mainittu traditio liittyy perinteiseen faktorianalyysiin, jota nykyisin kutsutaan eksploratiiviseksi faktorianalyysiksi (EFA). Siinä keskitytään korrelaatioiden kautta saadun muuttujien välisten yhteyksien tiedon käsittelyyn ja kuvailuun harvalukuisemmilla ulottuvuuksilla. 1960-luvulta lähtien on hiljalleen yleistynyt teoriapainotteisempi ja deduktiivisempi (teoriasta malli, jota testataan, data ei tuota mallia) SEM-tekniikka (Structural Equation Methodology). Sen eräs sovellusalue on konftrmatorinen faktorianalyysi (CFA). Meillä tulee olla aikaisemmasta tutkimuksesta ja teorianmuodostuksesta käsitys, millainen faktorirakenne olisi odotettavissa. Konfirmatorinen faktorianalyysi kertoo, kuinka hyvin aineisto tukee tätä hypoteesia ja antaa tuloksenaan suuntia sille, miten teoreettista mallia pitäisi trimmata ja muuttaa, jotta se paremmin vastaisi aineiston antamaa kuvaa. Saatu malli on edelleen uudella aineistolla testattava.

Eksploratiiviseen faktorianalyysiin ei juuri kuulu tilastollisen merkitsevyyden käyttäminen. Konfirmatorinen faktorianalyysi sisältää jo piiriinsä myös mahdollisuuden erilaisten tilastollisten hypoteesien ja yleistysten tekemiseen ja niiden hyvyyden arvioimiseen. Rajoitumme seuraavassa eksploratiiviseen faktorianalyysiin. Sen lähtöinformaatio on muuttujien välinen korrelaatiomatriisi. Muuttujat ovat kaikki samantasoisia (ei ole jakoa selittäjiin ja selityksen kohteisiin). Samoin syntyvät faktorit ovat samantasoisia. Ne eivät yleensä muodosta hierarkisia rakenteita. Muuttujajoukkoa ei jaeta enneusteisiin ja kriteereihin, vaikka tällainen jaottelu olisikin alustavasti tutkijan mielessä. Tämän vuoksi faktorianalyysia voidaan käyttää varsin vapaasti. Siihen liittyvä ongelmanasettelu voi olla hyvinkin löyhä: vaikkapa halu kartoittaa melko runsaan ja heterogeenisen osiojoukon takana olevia perusulottuvuuksia. Nuo ulottuvuudet saattaisivat olla sellaisenaan tutkimuksen tulos.

Tätä faktorianalyysin käytön löysyyttä on vuosien varrella voimakkaastikin kritikoitu. Asia on liitetty usein positivismin arvosteluun. Faktorianalyysin käyttö pitäisi liittyä tutkimuksen ongelmiin vastaamiseen. Perusteoksista ainakin klassikkoaseman saavuttanut Rummel (1970, Applied Factor Analysis) paneutuu faktorianalyysin perustehtäviin hyvinkin perusteellisesti. Faktorianalyysin käyttö tutkimuksessa tulee perustella tutkimustehtävästä ja tutkimusongelmista käsin.

Edellä olevan perusteella on jo helppo arvata, että faktorianalyysipohjaksi tarvitaan kaikkien mukana olevien muuttujien välinen korrelaatiomatriisi. Tässä yhteydessä on helpointa päästä eteenpäin, kun korrelaatioita kuvataan vektoreina. Mikä tahansa kahden muuttujan välinen korrelaatio voidaan esittää kahtena ykkösen pituisena vektorina, joiden välillä on tietty kulma. "Ykkösen pituinen" tarkoittaa sitä, että vektorit ajatellaan jaetuiksi asteikkoon, joka alkaa nollasta ja jatkuu yhteen asti. Vektoreiden välinen kulma riippuu korrelaatiosta siten, että nollakulma, siis vektoreiden ollessa päällekkäisiä, kuvaa ykkösen suuruista korrelaatiota, suora kulma edustaa nollakorrelaatiota ja 180 asteen kulma taas miinus yhden korrelaatiota. Näiden välille sijoittuvissa tapauksissa on korrelaatio vektorin projektio toisella, siis se "miten pitkänä se näkyy" toiselta vektorilta katsottuna. Seuraavat esimerkit selvittänevät tilanteen:



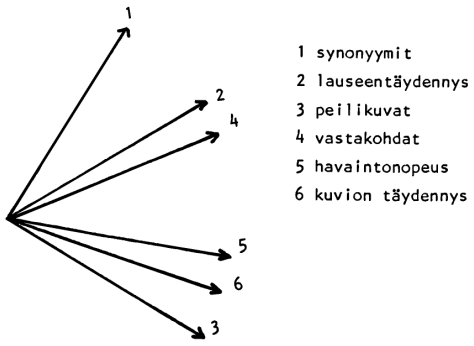
Koska vektorit ovat yhtä pitkiä, on tämä suhde symmetrinen: kummankin projektio toisella on sama.

Kuvitellaan nyt, että meidän tulisi ensimmäiseen kuvioon lisätä vektori, joka kuvaisi kolmatta muuttujaa, jolla on. 71 :n korrelaatio kumpaankin edelliseen. Nyt tuntuu tulevan vaikeuksia: jos yritämme sijoittaa vektorin kahden edellisen väliin, jäävät kulmat liian pieniksi. Jos taas sijoitamme sen edellisen kulman

ulkopuolelle, saamme kyllä aikaan oikean kulman toiseen vektoriin, mutta toiseen nähden joudumme suoraan kulmaan. Kuitenkin selvästikin voi olla olemassa kolmen muuttujan ryhmiä, joiden korrelaatiot ovat. 71. Ratkaisu tähän on kolmannen ulottuvuuden käyttö: emme enää pysykään paperin pinnalla, vaan nostamme vektorin irti edellisten vektorien keskiväliltä niin, että se tulee oikeaan kulmaan niihin nähden. Tietty korrelaatioiden joukko vaatii siis tietyn määrän ulottuvuuksia, jotta sen voisi kuvata täydellisesti). Nämä ulottuvuudet, tai niiden erilaiset arviot, ovat juuri faktoreita. Niiden määrä ei pysähdy esimerkissä olleeseen kolmeen; niitä voi olla enemmänkin, vaikkei tätä enää voi kunnolla piirroksin kuvata. Tavallinen tapa on käyttää suorakulmaista, ortogonaalista koordinaatistoa, mutta vinokulmaisiakin menetelmiä on käytettävissä. Pitäydymme tässä kuitenkin suorakulmaiseen tapaukseen.

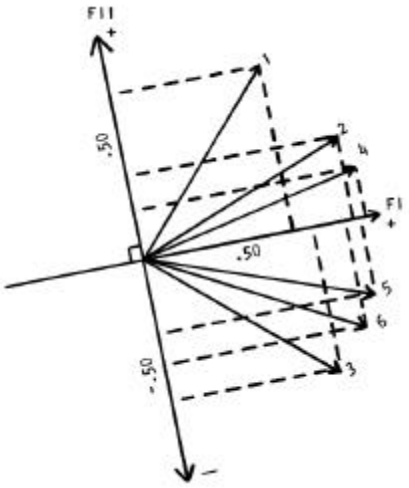
Havainnollisuuden vuoksi esitämme tässä kaksiulotteisen tapaukset mutta on muistettava, että se on yleensä todellisen tilanteen yksinkertaistus.

Kuvitellaanpa, että olemme hankkineet joltakin koehenkilöjoukolta tietoja kuudella eri muuttujalla, joiden välisistä korrelaatioista tulee seuraava vektorikuvaus:



Voimme todeta tästä mm, että muuttujien 2 ja 4 välillä on varsin korkea korrelaatio, muuttujien 1 ja 3 välillä on nollakorrelaatio jne. Faktorianalyysin tehtävänä on nyt kuvata tätä korrelaatioiden joukkoa mahdollisimman tehokkaasti ja vähin faktorein. Vektorikuvauksessa tämä merkitsee sellaisten uusien vektoreiden löytämistä, jotka asettuvat mahdollisimman lähelle mahdollisimman montaa entistä muuttujaa, jotta ne voisivat liikaa pakottamatta korvata niitä. Tähän ei ole mitään yhtä ja ainoaa menetelmää, mutta eräs tapa on asettaa ensimmäinen akseli siten, että se yksinään kuvaa mahdollisimman paljon kaikista vekto-

reista, selittää mahdollisimman paljon koko muuttujajoukon varianssista. Jos siis vektorikimpulla on jokin yleinen, yhteinen suunta, asetetaan ensimmäinen akseli, ts. ensimmäinen faktori, tähän suuntaan. Koska käsittelemme nyt suorakulmaista analyysiä, tulee toinen faktori automaattisesti suoraan kulmaan ensimmäiseen nähden. Se voidaan kuitenkin asettaa parhaiten selittävään asentoon sillä tasolla, jonka kaikki mahdolliset suorat kulmat muodostavat. Meidän piirretyssä tasolla kuvattavassa tapauksessamme se kuitenkin saa vain yhden mahdollisen ratkaisun. Seuraava faktori tulisi jälleen suoraan kulmaan edellisiin nähden jne. Kuten äsken tutkimme muuttujien korrelaatioita toisiinsa vektorien projektioina, voimme nyt tutkia muuttujien korrelaatioita uusiin akseleihin, faktoreihin. Nämä korrelaatiot, faktorilataukset, ovat analyysin keskeisin tulos. Kukin muuttuja saa jonkin latauksen jokaisella faktorilla. Voimme piirtää tilanteen ja arvioida lataukset kahdella faktorilla:



Kun nyt ilmoitamme kunkin muuttujan lataukset kullakin faktorilla, saamme faktorimatriisin, joka tässä tapauksessa näyttää suunnilleen seuraavalta:

muuttuja	F I	F II
1 synonyymit	.65	.75
2 lausecntäydennys	.90	.35
3 peilikuvat	.70	—.65
4 vastakohdat	.95	.20
5 havaintonopeus	.92	—.35
6 kuvion täydennys	.85	—.48

Latausten suuruudet ilmoittavat, "kuinka paljon tekemistä muuttujalla on faktorin kanssa" ja faktoreille annetaan tulkinta sen mukaan, minkä tyyppiset muuttujat sille voimakkaimmin latautuvat. Analyysin tekotavan mukaan on kuitenkin varsinkin ensimmäinen faktori sikäli epämääräinen, että kaikilla muuttujilla on sillä kohtalainen tai korkea lataus (tekotapansakin perusteella). Usein tämä on haitaksi, mutta välillä se on täysin mielekästä ja tällainen komponentti voidaan sellaisenaan tulkita. Esimerkiksi tässä tapauksessa voimme pitää ensimmäistä faktoria jonkinlaisen yleisen älykkyyden edustajana. Tilanne on sikäli tyypillinen, että varsinkin kykytestejä faktoroitaessa löytyy usein yleinen faktori, jolla on positiiviset lataukset kaikkien muuttujien kohdalla.

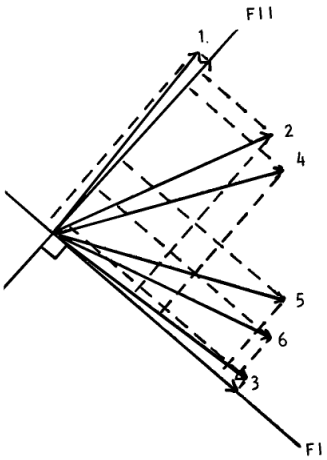
Toinen faktori on myös tyypillinen sikäli, että se on kaksipäinen, bipolaarinen, ts. sillä on selviä sekä positiivisia että negatiivisia latauksia. Tällaisessa muuttujajoukossa tämä on luonnollista, koska ensimmäinen, koko joukkoa mahdollisimman hyvin kuvaava faktori asetettiin vektoreiden keskelle, joka tietyllä tavalla tuli vakioitua. Toisen faktorin kannalta joukko siis tuli jaetuksi nollapisteen molemmille puolille, positiiviseen ja negatiiviseen päähän.

Tämä faktori on sikäli helppo tulkita, että kaikki kielelliset testit tulivat sen positiiviseen päähän ja kaikki kuvalliset testit taas saivat negatiivisen etumerkin. On huomattava, että sekä negatiiviset että positiiviset lataukset ovat yhtä paljon huomioonotettavia, ne vain sijaitsevat dimension eri päissä. Vain lähellä nollaa olevat lataukset ovat ko. faktorin kannalta merkityksettömiä tai vähän painottuvia. Eri päihin tulevat muuttujat ovat tavallaan toistensa vastakohtia, ne korreloivat negatiivisesti. Tämä ei kuitenkaan välttämättä näy alkuperäisessä korrelaatiomatriisissa, koska jotkin muut tekijät voivat peittää sen näkyvistä. Niinpä meidän tapauksessamme korreloivat verbaaliset ja kuvalliset testit negatiivisesti (toisella faktorilla), kun yleisen älykkyyden osuus (ensimmäinen faktori) on poistettu. Näinhän totesimme jo osittaiskorrelaation yhteydessä asian usein olevankin. Toinen komponentti nayttää siis muodostavan verbaalinen-spatiaalinen -dimension, jonka ääripäitä edustavat synonyymi- ja peilikuvatesti.

Tässä vaiheessa voidaan huomata, että mahdollisia asentoja faktoreille (koordinaatistolle) on useita, itse asiassa lukematon määrä. Muuttujien vektorikuvion pysyessä paikallaan, voimme kiertää faktorien muodostamaa X- Y -koordinaatistoa mielivaltaisesti ja jokainen asento yhtä hyvin reprodusoi muuttujien väliset korrelaatiot (ns. faktorianalyysin peruskaava). Faktorianalyysin lähtö-

kohta on korrelaatiomatriisi. Mitään muuta tietoa ei tarvita. Kun faktorianalyysin laskentatekniikka asettaa siihen ensimmäisen pääkomponentin tai pääakselin, laitetaan se kulkemaan kohdasta, jossa se kerää kaikista muuttujista keskimäärin suurimman määrän muuttujavektorien projektioista. Maksimoidaan ensimmäisen pääkomponentin selitysosuus. Toinen pääkomponentti asetetaan kulkemaan siten, että se jäännöksestä selittää mahdollisimman paljon (ehdolla että se on korreloimaton eli suorassa kulmassa jo viritettyyn ensimmäiseen pääkomponenttiin nähden). Kun edellä tulkitsimme tällaisten faktoreiden sisältöä, tulimme tehneeksi aika lailla virheellisiä päätelmiä. Alkuratkaisun tulkinta yleensä ei olekaan mielekäs. Faktorit voidaan siis sijoittaa oikeastaan miten vaan, kunhan niiden välillä pysyvä kulma pidetään suorana ja faktorien alkupiste on sama kuin muuttujavektoreiden.

Tällaisessa tilanteessa alkuratkaisua (joka on matemaattisin perustein laadittu) muutetaan. Faktorikoordinaatistoa kierretään siten, että se vastaa teoreettista asetelmaa parhaiten tai muutoin on tulkinnallisesti selkein ja mielekkäin. Vaikka alkuratkaisulla on oma matemaattinen perustansa, jatketaan yleensä rotatoimalla alkuperäistä koordinaatistoa. Etsitään tulkinnallisesta parasta rotatoitua faktorimatriisia. Kiertämällä koordinaatistoa myötäpäivään löydämme sille asennon, jossa muuttuja Xl on aika puhdas F2:n edustaja ja muuttuja X3 verraten puhdas F1:n edustaja.



Rotatointi tarkoittaa sananmukaisesti kiertämistä; tässä tapauksessa kierretään faktorien muodostamaa koordinaatistoa muuttujajoukossa siten, että saadaan jonkin kriteerin mukaan paras ratkaisu. Eräs kriteeri parhaalle ratkaisulle on "yksinkertainen rakenne" (simple structure), jossa kukin muuttuja latautuu sel-

västi vain yhdellä faktorilla. Epämääräisiä tilanteita, joissa sama muuttuja näyttäisi kuuluvan yhtä lailla usealle faktorilIe, pyritään välttämään. Yksinkertaiseen rakenteeseenkin voidaan pyrkiä monella tavalla; eräs lähelle tätä ideaalia pääsevä menetelmä on maksimoida latausten neliöiden varianssit kullakin faktorilla ("varimax"). Siinä siis pyritään saamaan samalle faktorille mahdollisimman suuria ja mahdollisimman pieniä latauksia, jolloin niiden (latausten faktorimatriisin sarakkeilla) vaihtelu, varianssi, on niin suuri kuin mahdollista. Selitysosuuksien summa pysyy ennallaan rotatoinnissa. Selitysosuuden jakautuminen faktoreille muuttuu. Yleensä (mutta ei välttämättä) faktoreiden selitysosuuksien järjestys säilyy samana.

Meidän yksinkertaisessa esimerkissämme on jo silmämääräisestikin tarkastellen kaksi toisistaan erotettavissa olevaa muuttujakimppua, toisaalta 1, 2 ja 4 sekä toisaalta 5, 6 ja 3. Luonnollisimmalta rotaatiolta tuntuisi tällöin yrittää saada kumpikin kimppu mahdollisimman selkeästi omalle faktorilleen. Esitämme edellisellä sivulla äskeisen vektorikimpun, jossa faktoreita on kierretty mahdollisimman yksinkertaisen rakenteen aikaansaamiseksi.
Nyt saamme suunnilleen seuraavassa esitetyn faktorimatriisin. Matriisiin on lisätty kommunaliteetit ja ominaisarvot, joiden merkitys täsmennetään myöhemmin. Kuusi muuttujaa virittää alun perin kuusiulotteisen avaruuden. Siitä voitaisin eristää 6 korreloimatonta pääkomponenttia (mikä ei olisi kovin taloudellista). Tulemme toimeen varsin hyvin kahdella ensimmäisellä, jotka tyhjentävät melko hyvin käytettävissä olevan informaation.

muuttuja	F I	F II	h^2
1 synonyymit	—.05	.98	.96
2 lauseentäydennys	.40	.90	.97
3 peilikuvat	.98	.05	.96
4 vastakohdat	.55	.80	.94
5 havaintonopeus	.85	.35	.85
6 kuvion täydennys	.95	.25	.97
ominaisarvot	3.05	2.60	5.65

Aivan yksinkertaiseen rakenteeseen ei tässä muuttujajoukossa päästä. Jo vektorikuviota tarkastelemalla on helppo havaita, etteivät esim. muuttujat 2 ja 4 kuulu täysin yksiselitteisesti kumpaankaan joukkoon. Selvää painottumista on kuitenkin siten, että peilikuvat, havaintonopeus ja kuvion täydennys kuuluvat ensimmäiselle faktorille ja synonyymit, lauseentäydennys ja vastakohdat toiselle.

Ensimmäinen faktori on lähes identtinen peilikuvatestin kanssa ja toinen on hyvin lähellä synonyymitestiä. Faktorit ovat melko selvästi nimitettävissä spatiaaliseksi ja verbaaliseksi tekijäksi (faktoriksi). Verrattuna aikaisempaan analyysiin on näiden ero täsmentynyt (ne ovat eri faktoreilla) ja yleisen älykkyyden osuus tulkinnassa on hävinnyt.

Kommunaliteetit ja ominaisarvot ovat molemmat toiseen korotettujen latausten summia, kommunaliteetit rivisuuntaan ja ominaisarvot sarakesuuntaan. Koska lataukset ovat itse asiassa korrelaatioita (muuttujan ja faktorin välillä) olemme jälleen tekemisissä toiseen korotettujen korrelaatiokertoimien kanssa. Nämähän, kuten muistettaneen, olivat selitysosuuksia, ne ilmaisevat kuinka paljon muuttuja selittää toisesta. Niin on tässäkin: kun lataukset korotetaan toiseen saadaan selville, kuinka paljon faktori selittää kustakin muuttujasta (tai päinvastoin). Näiden rivisuuntaiset summat, kommunaliteetit, siis ilmaisevat kuinka paljon faktorit yhteensä, siis koko analyysi, selittävät ko. muuttujasta. Synonyymitestin varianssista siis on esimerkissä selitetty 96.3 %, joka tulee lähes yksinomaan toiselta faktorilta. Tämän esimerkin keinotekoisesta kaksidimensionaalisuudesta johtuen saadaan selitetyksi lähes kaikki vaihtelu, mitä muuttujissa on: jos jaamme kommunaliteettien summan muuttujien määrällä, 5.649:6, saamme tietää, että muuttujien vaihtelusta on selitetty keskimäärin 94 %. Se, miten tämä jakaantuu faktoreittain, tulee ominaisarvoista vastaavasti: ensimmäinen faktori selittää 3.051:6 = 51 % ja toinen 2.598:6 = 43 %.

Vaikka tässä aineistossa ei enää kahden faktorin jälkeen jäänyt juuri mitään vaihtelua jäljelle, voidaan faktorointia (ortogonaalisiin komponentteihin jakamista) siis periaatteessa jatkaa ja saada aineistoon uusia näkökulmia. Tuotettujen (rotaatioon mukaan otettujen) faktoreiden määrä on usein jossakin määrin subjektiivinen asia; analyysin tekijän on valittava mielestään parhaat kriteerit sille, mikä on sopiva määrä faktoreita. Faktorianalyysin ohjelmissa voi yleensä valita vapaasti faktoreiden lukumäärän. On myös eräitä perinteisesti käytettyjä peukalosääntöjä. Tavallisesti tuotetaan kuitenkin useita analyysejä, joissa on eri määriä faktoreita ja näistä valitaan "paras". Selvää on, että selittävyys on eräs kriteeri: on turha jatkaa, kun uudet faktorit selittävät niin vähän, että se alkaa olla sattumaa. Ehkä tärkein kriteeri on kuitenkin mielekkyys. Ei ole järkevää tuottaa sellaisia faktoreita, joita ei pysty millään uskottavalla tavalla enää tulkitsemaan.

Faktorianalyysin tekeminen on ollut jo liki neljän vuosikymmenen ajan hyvin helppo toimenpide. Tutkijan tarvitsee vain osoittaa ne muuttujat, jotka otetaan analyysiin mukaan. Mitään jakoa selittäviin ja selitettäviin ei faktorianalyysi

tee. Lähtötietonaan se tarvitsee muuttujien välisen korrelaatiotaulun (korrelaatiomatriisin). Pahimmillaan korrelaatiot voivat olla lasketut kvalitatiivisista moniluokkaisista muuttujista. Varmaan onnistuu sekin, että luotaisiin 50 kappaletta satunnaismuuttujia ja kuhunkin arvottaisiin esim. 400 peräkkäistä satunnaisarvoa (N=400). Muuttujat korreloitaisiin. Pääkomponentteja syntyisi ja muutama ensimmäinen voitaisiin rotatoida faktoreiksi. Tulkintakin usein on mahdollista (eli tutkija projisioi huuhaa-numeroihin näkemiään/toivomiaan asioita). Sanallisesti lahjakkaat tutkijat ovat hyvin taitavia konstruoimaan sisältöä sattumatilanteeseenkin (jos/kun eivät tiedä että kaikki onkin sattumaa).

Faktorit ovat myös saman arvoisia. Eksploratiivinen faktorianalyysi (EFA) ei pysty käsittelemään hierarkkisia rakenteita. Faktorien korreloimattomuus on vaatimus, josta voidaan kuitenkin luopua rotaatiovaiheessa. Rotaatiovaiheen menettelyt ovat siis joko suorakulmaisia (ortogonaalisia) tai vinoja (engl. oblique). Gorsuch'n klassikkoteos esittelee yhteensä 19 erilaista rotaatio- ratkaisua (kutsuen esittelemäänsä näytteeksi, useampiakin menettelyjä on olemassa). Koska itse faktorintimenettelyjäkin (pääkomponentti, pääakseli, suurimman uskottavuuden jne) on käytettävissä kymmenkunta, niin kirjo faktorianalyyseissaon erittäin suuri. Faktorianalyysi ei ole yksi yhtenäinen menetelmä.

Kun kaikista aineistosta (myös satunnaislukutauluista) syntyy jonkinlainen faktorirakenne, on tämä johtanut vuosien aikana moniin täysin perustelemattomiin analyysimallin sovelluksiin. Eksploratiivinen faktorianalyysi on saanut huonon, "positivistisen" maineen. Teorian, ongelmien ja data-analyysin täytyy muodostaa sykli, jossa asiat liittyvät perustellulla tavalla toisiinsa. Vasta tällöin data-analyysilla on mielekästä annettavaa.

Muuttujien tulee olla 3- tai useampiluokkaisia kvantitatiivisia muuttujia. Kun muuttujan koodi kasvaa pitää sen indikoiman ominaisuudenkin voida ajatella monotonisesti kasvavan. Pseudo-intervallisuus on siis jonkinlainen perusedellytys. Jakaumien muodot vaikuttavat korrelaatioihin vahvasti. Sen vuoksi analyysin tulee alkaa aineiston alustavalla läpi käymisellä perusasioiden osalta (muuttujien jakaumat ja korrelaatiot). Eräänlainen rajatapaus käytölle on kaksiarvoinen (dikotominen) muuttuja: esim. sukupuoli tai yksin/parisuhteessa eläminen tmv. Sen tulomomenttikerrointa jatkuvaan muuttujaan nimitetään pistebiseriaaliseksi korrelaatioksi. Analyysissa voi olla jokunen tällainenkin muuttuja. Jos joku muuttuja ei saa vaihtuvia arvoja (onkin vakio) tai mukana ovat

muuttujat x1 ... x10 ja niiden summa x 11, syntyy tilanne joka tuottaa korrelaatiomatriisin, jonka tekniset ominaisuudet eivät ole riittävät. Joskus muuttujat (etenkin indeksiluvut tai summamuuttujat) sisältävät samoja osatekijöitä. Muuttujien välillä on tällöin teknistä korrelaatiota. Tätä ei analyysiohjelma huomaa. Tekijän tulee olla selvillä korrelaatiokertoimesta ja sen suuruuteen vaikuttavista asioista.

Nyt teemme kirjamme aineistosta pienen faktorianalyysin ja etenemme siitä regressioanalyysiin. Tässä korostuvat tekniset asiat. Samoin analyysi tehdään perinteisellä tavalla: muuttujille asetetaan kommunaliteettiarvot. Faktoripisteitä tehdään usealla tavalla: sekä summamuuttujina että regressioestimoituina faktoripisteinä.

Esimerkkimme rajoitamme minimiin. Meillä on neljä kykyä indikoivaa muuttujaa x3 ...x6. Korrelaatiomatriisista alkusivuilta näemme niiden korrelaatiot ja näemme, että ne kuvautuvat hyvin kaksiulotteisesti. Muuttujaa x7 käytämme selityksen kohteena regressioanalyysissa. Selvitämme ensin, olisiko järkevää koota selittäviä muuttujia jotenkin muuttujaryhmiksi ja käyttää niitä alkuperäisten muuttujien asemasta.

Aluksi kuvaavaa perustietoa (jota faktorianalyysin tekninen suoritus ei edes tarvitse). Joskus törmää ajatukseen, että korkean faktorilatauksen saanut muuttuja on myös saanut korkean keskiarvon tmv. Rakenne ja taso ovat kaksi eri asiaa:

Muuttuja	Keskiarvo	Hajonta	N
V3	25.47	2.57	30
V4	34.97	1.83	30
V5	79.73	10.39	30
V6	7.23	1.10	30

Muuttujien väliset korrelaatiot ovat:

	V3	V4	V5	V6
V3	1.000	.246	.909	.143
V4	.246	1.000	.079	.738
V5	.909	.079	1.000	-.015
V6	.143	.738	-.015	1.000

Korrelaatioista nähdään näin yksinkertaisessa tapauksessa suoraan muuttujaryppäät v3, v5 ja v4ja v6 sekä näiden melko selvä riippumattomuus toisistaan.

Kaiser-Meyer-Olkin -indeksi (MSA) ja Bartlettin khii-toiseen testi saavat jäädä väliin tulostuksesta tässä vaiheessa.

Perinteinen faktorianalyysi kohdistaa faktoroinnin ns. redusoituun korrelaatiomatriisiin. Siinä diagonaalin ykköset korvataan saatavissa olevalla hyvällä yhteisen osuuden alkuarvolla. Tällaiseksi voidaan osoittaa kyseisen muuttujan multippelikorrelaation neliö muihin muuttujiin. Faktoroinnin yhteydessä saatu uusi arvio asetetaan uudeksi alkuarvoksi ja uusitaan faktorien eristäminen pääakselimenetelmällä. Tätä iteroinniksi kutsuttua menettelyä toistetaan, kunnes diagonaalin arvot vakiintuvat. Alkuarvojen ja lopullisten arvojen vertailua kannattaa suorittaa. Suuret erot toimivat diagnostiikkana faktorianalyysissa. Meidän tapauksessamme alkuarvot kasvavat, mutta mitään hälyttävää ei ilmene. Huono alku- ja/tai loppuarvo panisi miettimään, mitä kyseiselle muuttujalle pitäisi tehdä (esim. poistaa analyysista).

	Kommunaliteetit	
	Alkuarvo	Lopullinen
V3	.858	.934
V4	.576	.757
V5	.852	.920
V6	.556	.729

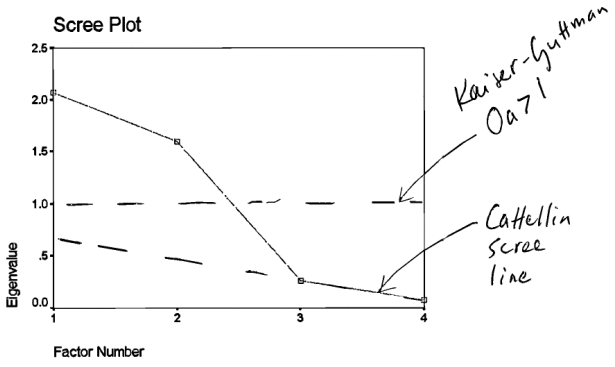
Yleinen on myös tapa laittaa diagonaalille ykköset ja ottaa muutamia ensimmäisiä pääkomponentteja rotaatioon. Pääkomponenttifaktoroinnin (PC) ja pääakselifaktoroinnin (PAF) ero onkin vain tässä. Yleensä pääkomponenttien faktoroinnissa saadaan korkeampi lataustaso ja menettelyllä on myös taipumus tuottaa suurempi määrä faktoreita, joista osa perustuu vain hyvin pieneen muuttujamäärään. Ero on kumminkin sekä teoreettisesti että käytännöllisesti tärkeä, vaikka lopputulos usein on melko saman kaltainen.

Pääkomponentit lasketaan pääakselifaktoroinnin yhteydessäkin. Ominai - sarvot ennen rotaatiota ja niiden arvoista piirretty käyrä (Cattellin scree) perustuvat rotatoimattomiin pääkomponentteihin, joita pitää tulla yhtä monta kuin on muuttujiakin.

Faktori	Ominaisarvot (pca)			Ominaisarvot (pfa)			Ominaisarvot (rotatoitu)		
	Oa	Sel-%	Cum-%	Oa	Sel-%	Cum-%	Oa	Sel-%	Cum-%
1	2.073	51.83	51.833	1.954	48.839	48.839	1.839	45.972	45.972
2	1.596	39.90	91.736	1.387	34.667	83.506	1.501	37.533	83.506
3	.256	6.40	98.132						
4	.075	1.87	100.000						

Viimeinenkin pca-ominaisarvo on positiivinen, joten korrelaatiomatriisi on täyttä astetta. Kommunaliteettien asettamisen ja iteroinnin jälkeen selitysprosentiksi tulee 83.5, joka (rotatoitujen) tulkittujen faktoreiden osalta on: ensimmäinen 46.0 % ja toinen 37.5 %.

Pääkomponenttien ominaisarvojen (Cattellin scree-testiksi nimetty) käyrä näyttää seuraavalta:



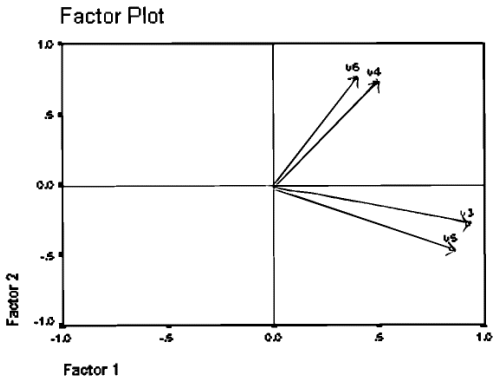
Johtopäätös on selvä. Kahden faktorin jälkeen ei enää oleellista varianssia ole jäljellä. Kuvio on siis pääkomponenteista. Rotatoimaton faktorimatriisi jätetään esittämättä. Varimax-rotaation (menettely pyrkii maksimoimaan lataussarakkeiden varianssin) jälkeen saadaan seuraava faktorilatausmatriisi. Latauksen voi tulkita muuttujan ja faktorin korrelaatioksi.

Rotatoitu faktorimatriisi

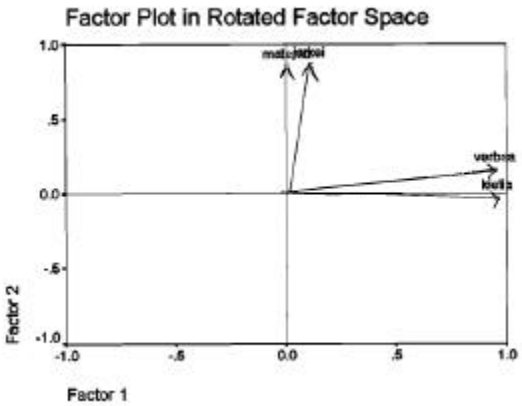
	Faktorit 1	2
V3	.953	.162
V4	.109	.863
V5	.959	-.028
V6	.007	.854

Tulkinnassa ei liene mitään ongelmaa. Muuttujat kuuluvat varsin selvästi toisaalta kielelliseen ja toisaalta matemaattis-järkeilylliseen faktoriin.

Rotatoimaton tilanne:



Rotatoitu tilanne:



Kun näillä kahdella tekijällä pyrimme selittämään opintomenestystä, meidän on päätettävä miten uudet muuttujat (2 kpl) muodostamme. Summamuuttujien muodostus näyttää konkreettiselta ja selkeältä. Sitä se onkin, kunhan muistamme, että yhteen laskettavilla muuttujilla tulee olla likipitäen sama hajonta. Jos muuttujat olisivat kaikki 5-portaisella asteikolla, tämä seikka ei ehkä merkitsisi juuri mitään. Hajonnat olisivat riittävän samankaltaisia. Nyt tilanne ei ole niin hyvä (kuten muuttujien hajontatiedoista voimme yllä todeta). Meidän tulee ensin standardoida muuttujat. Sen jälkeen teemme kaksi summamuuttujaa ja niillä selitämme opintomenestystä.

Faktoripisteiksi nimitetään taas pistemääriä, jotka on estimoitu regressioanalyysiä hyväksi käyttäen. Faktorianalyysistä saamme latauksista tiedon kuinka muuttujat korreloivat kunkin faktorin kanssa. Korrelaatiomatriisissa on taas tieto kuinka muuttujat korreloivat keskenään. Ennustamme siis kumpaakin faktoria neljällä muuttujalla. Saatuja painokertoimia käyttäen muuttujat painottuvat summaan. Summa tulee ominaisuuksiltaan sellaiseksi, että sen keskiarvo on nolla ja hajonta regressiomenettelyssä saatu multippelikorrelaatio (varianssi=R-toiseen). Saadut estimaatit perustuvat siis niihinkin muuttujiin, jotka eivät ole juuri tätä tiettyä faktoria latauksellaan määrittämässä.

Summapisteet ja faktoripisteet yleensä korreloivat melko korkeasti (yleensä yli .8). Summapisteille voidaan laskea Cronbachin alfa reliabiliteetin arvioimiseksi. Faktoroinnille, faktoreille ja faktoripisteille (regressioestimoiduille) on puolestaan varsin hankala tuottaa reliaabeliusarvioita. (ks. kuitenkin Tarkkonen, samoin Vehkalahti, Survo-ohjelma).

Summapisteiden keskinäisistä korrelaatioista heijastuu faktorien edustamien käsitteiden korreloituminen. Faktoripisteetkin korreloivat, mutta lähtökohtaisesti ne pakotetaan olemaan mahdollisimman riippumattomat (suorakulmaisessa rotaatiossa). Ja huomattava on, että pääkomponenttianalyysin tapaan (eli ykköset diagonaalilla) tehdyssä faktorianalyysissä faktoripisteet ovat aina täysin korreloimattomat ja niiden hajonta on 1 (eli niitä ei estimoida, ne voidaan laskea tarkasti) mikä joskus harhaannuttaa ajattelemaan ilmiöstä liian yksioikoisesti.

Yleiseksi menettelytavaksi voidaan suosittaa sekä varimax- että promax-rotaatioiden suoritusta ja vertailua. Promaxin lähtötilanne on varimax, jota se pyrkii parantamaan yksinkertaista rakennetta luopumalla faktorien nollakorrelaatioista, eli se on vinorotaatio. Summamuuttujien korrelaatiot keskenään on myös hyvä laskea ja raportoida.

Regressioanalyysit teemme viidellä eri tavalla, joiden keskeiset tulokset ovat seuraavassa:

```
Riippuva muuttuja v7 (opintomenestys)

Muuttuja    beta  p-arvo    r1     r2   omaosuus

V3          .075   .769     .812   .028   .001
V4          .048   .746     .316   .031   .001
V5          .758   .005     .826   .292   .048
V6          .257   .082     .291   .171   .029
R-toiseen = .776, p = .000 (F = 21.71, df1 = 4, df2 = 25)

V3+V5       .814   .000     .835   .812   .639
V4+V6       .260   .010     .328   .260   .068
R-toiseen = .765, p = .000 (F = 43.91, df1 = 2, df2 = 27)

Z3+Z5       .810   .000     .838   .804   .646
Z4+Z6       .225   .027     .326   .223   .050
R-toiseen = .752, p = .000 (F = 41.03, df1 = 2, df2 = 27)

FPPAF1      .817   .000     .821   .817   .667
FPPAF2      .260   .013     .272   .260   .068
R-toiseen = .742, p = .000 (F = 38.82, df1 = 2, df2 = 27)

FPPCA1      .821   .000     .821   .821   .674
FPPCA2      .273   .009     .273   .273   .075
R-toiseen = .749, p = .000 (F = 40.20, df1 = 2, df2 = 27

P-arvo on betan, r2:n ja omaosuuden tilastollisen merkitse-
vyystestauksen tulos (tietoja siitä kirjan jälkimmäisessä
osassa

r1 on muuttujan suora (nolla-asteen) r kriteeriin
r2 on muuttujan semipartiaalikorrelaatio kriteeriin
omaosuus on r2 toiseen korotettuna
```

Käyttämällä kaikkia neljää muuttujaa selittävinä muuttujina olisimme saaneet aikaan "parhaimman" selityksen (korkein R-toiseen). Multikollineaarisuuksista johtuen se olisi kumminkin ollut yksityisten muuttujien osalta sekava (osakorrelaatiot ja niiden neliöt = omaosuudet). Samoin kun P/N -suhde kasvaa (eli df1/df2 -suhde), niin regressioanalyysin yhteydessä saadaan liian optimistisia tuloksia. Ns. shrinkage-korjaus on yritys poistaa tämä harha. Ilmiön selittäminen ottaa tilaa, joten jää myöhempiin opintoihin.

Toinen tilanne vastaa kirjan edellistä painosta. Siinä on kauneusvirheenä se, että melko erilaisen hajonnan omaavia muuttujia on suoraan laskettu summamuuttujiksi.

Kolmannessa tilanteessa summa on tehty Z-pistemääriksi muunnetuista muuttujista.

Neljäs tilanne on faktorianalyysin regressioestimoiduista faktoripisteistä.

Viides tilanne on faktoripisteet pääkomponenttianalyyista, jossa kaksi pääkomponenttia on rotatoitu. Lasketut faktoripisteet ovat täysin korreloimattomia, mikä näkyy myös tuloksista.

Näin selvässä tilanteessa vain ensimmäinen tapa on selvästi ongelmallinen. Selittävät muuttujat korreloivat liikaa keskenään. Niitä pitää siis yhdistellä sopivalla tavalla.

Tässä kahdella yleisemmällä muuttujalla olemme yksinkertaistaneet tilannetta, toimineet primaarimuuttujia luotettavammilla selittävillä muuttujilla ja pitäneet P/N -suhteen pienenä. Tulos on helpompi raportoidakin. Opintomenestys selittyy kielellisillä tekijöillä huomattavasti paremmin saadun tuloksen valossa.

Regressioanalyysin yhteydessä tehdään aina ns. regressiodiagnostiikka, jolla selvitetään mitkä regressioanalyysin perusedellytyksistä mahdollisesti täyttyvät huonosti. Tätä ei tässä tarkastella pidemmälle. Tärkein asia lienee, että residuaalitarkastelussa ei löydy pahoja poikkeamia normaalista jakautumisesta.

Kirjan jälkiosaa ennakoiden mukana on jo tilastollista merkitsevyyttä kuvaavia tunnuslukuja.

Eksploratiivisen faktorianalyysin yhteydessä ei juuri käytetä merkitsevyystestauksen apua. Konfirmatorinen faktorianalyysi taas puolestaan yhdistää sekä kuvauksen että päättelyn. Validiuden ja reliaabeliuden käsitteitä ei tässä ryhdytä pohtimaan.

Harjoitustyönä voit tehdä tällaisen regressioanalyysin Spss-ohjelmalla. Standardoiminen, summamuuttujan laatiminen ja regressioanalyysin tulostus tulevat tutuiksi.

Muutama faktorianalyysin asia:

- rotatoidussa ratkaisussa selitysosuudet eivät välttämättä ole laskevassa järjestyksessä

- jos muutat 5-portaisen muuttujajoukon pisteityksen 1 ...5 käänteiseksi 5... 1 jokaisen muuttujan osalta, saatu korrelaatiomatriisi on entinen samoin fakto-

rointi, mutta summamuuttujat ja/tai faktoripisteet on käännetty (reflektoitu). Kun käännetään vain muutama muuttuja, on syytä olla tarkkana, että tulkinnat yms. pysyvät oikeina.

- vinorotaatiossa (direct oblimin) faktorit voivat muuttaa paikkaansa (kuinka mones faktori) ja/tai reflektoitua (kääntyä). Faktorin reflektio on faktorivektorin kääntyminen 180 astetta eli lataussarake on kerrottu -1:llä.

- summamuuttujan muodostamista varten osioiden pisteitys pitää olla "saman suuntainen", hajontaerot eivät saa olla suuria. Mikä on suuri? Vaikkapa se jos pienimmän ja suurimman hajonnan suhde on .5 noin peukalosääntönä.

- on syytä eri tavoin varmistaa, että summamuuttujat ja/tai faktoripisteet edustavat faktoreita tarkoitetun suuntaisesti.

Nunnallyn klassikkoteoksessa (Psychometric Theory) on mukava pikku luku: How to fool yourself with factor analysis. Suositellaan.

Sallittaneen tilaa myös seuraaville graafisille esityksille. Ensimmäiset liittyvät regressioanalyysiin. Pohjana on se, että muutama esitystapa on mahdollista tehdä itselle ja ymmärtää tilanteessa, jossa meillä on kaksi ennuste- tai selittävää muuttujaa ja yksi kriteeri. Takana on korrelaatiokertoimen kolme esitystapaa. (1) Muuttujat voidaan kuvata yksikkövektoreina N-ulotteisessa havaintoyksikköavaruudessa. (2) Havaintoyksiköt voidaan kuvata pisteinä muuttujien muodostamassa ortogonaalikoordinaatistossa. (3) Muuttujien välisiä yhteyksiä voidaan kohtuullisen virheettömästi kuvata selitysosuuksina Vennin diagrammiesityksenä. Tarkemmin asiasta tämän luvun lopussa.

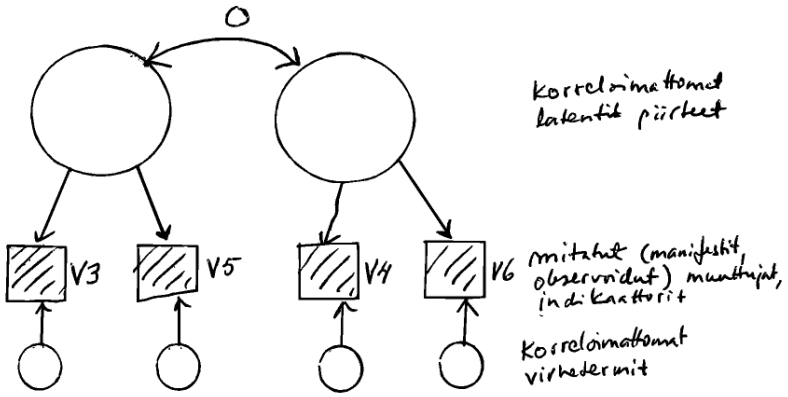
Multikollineaarisuuden (selittävien muuttujien korkeahkot korrelaatiot keskenään) aiheuttamat latausten etumerkkikysymykset on erityisen hyvin ymmärrettävissä muuttujien vektoriesityksestä.

Sekä regressioanalyysi että faktorianalyysi operoivat havaittujen muuttujien välisillä korrelaatioilla suoraan, olettamatta mittauksiin sisältyvän erityistä reliabiliteetin puutetta. Psykometriikan alkeista tiedämme, että havaittujen korrelaatioiden takana olevat tosiarvojen väliset korrelaatiot ovat korkeampia, koska reliabiliteetin puute attenuoi ("laimentaa") niitä. Vuosikymmenien myötä alkaa SEM-tekniikoiden puitteissa olevan mahdollista ottaa huomioon indikaattorien

reliabiliteetit ja yhdistää mittaus- ja rakennemallit yhtäaikaisen estimoinnin piiriin. Datalle asetetut vaatimukset ja teoriavetoisuus asettavat omat vaatimuksensa.

Regressioanalyysin ja (eksploratiivisen) faktorianalyysin voi ymmärtää alkeelliseksi mallintamiseksi.

Esimerkissämme tunnemme neljän muuttujan väliset korrelaatiot. Kuinka hyvin nuo korrelaatiot ovat reprodusoitavissa sellaisella mallilla, jossa on kaksi korreloimatonta faktoria. On huomattava, että rotatoimaton ja erilaiset kuvausavaruuden rotaatiot "palauttavat" (reprodusoivat) alkuperäiset korrelaatiot yhtä hyvin. Tätä yhtälöä (eli rivien välisten latausten tulojen summaa) nimitetään joskus faktorianalyysin perusyhtälöksi. Rotaatioratkaisun valinta ja tulkinnan suorittaminen ovatkin mallinnuksen kannalta ongelmallisia ja mielivaltaisiakin asioita. Mikäli lähtisimme oheisesta mallista ja testaisimme sen kykyä kuvata dataamme, tekisimme konfirmatorista faktorianalyysia (jota Spss-ohjelmassa ei ole). Esimerkkimme on myös valitettavasti testaus kelvoton ei-identifioituvuutensa vuoksi: liian vähän datapisteitä estimoitavaa parametria kohden. Mutta idea on tämä: testaisimme onko alkuperäisten korrelaatioiden riittävän hyvä kuvaus seuraavassa:



Eräät oppikirjat kehottavat (mm. Yli-Luoman Spss-opas ja Nummenmaa & al.) tekemään faktoroinnin siten, että se suoritus aloitetaan laittamalla kaikki kyseisen alueen muuttujat reliabiliteetin (muuttujajoukon homogeenisuuden) tarkasteluun (Cronbachin alfa). Negatiiviset (eri suuntaan mittaavat) osiot käännetään tässä vaiheessa samaan suuntaan muiden kanssa. Alustava karsinta suoritetaan

sen perusteella lisääkö osion mukana olo reliaabeliuden indeksiä ja millainen sen korjattu/puhdistettu osiokorrelaatio on.

Tämän kirjan ohje ensimmäiseksi askeleeksi ei ole tämä. Osioiden kuulumisen yhteiseen joukkoon voidaan havaitanäet ehkä paremmin muulla tavalla. Alfan laskeminen myös pakottaa eräiden muuttujien liian varhain suoritettavaan ja turhaan pisteityksen kääntämiseen. Kun faktoreiden perusteella muodostetaan summamuuttujia, on (faktorilatauksen) etumerkki otettava huomioon. Osioiden on mitattava kohdetta saman suuntaisesti. On selkeämpi suorittaa faktorianalyysi alkuperäisillä muuttujapisteityksillä ja vasta sitten esim. summamuuttujia ja reliabiliteetteja varten kääntää muuttujapisteityksiä.

Seuraava ohje lienee varsin suositeltava:

1) Samaan alueeseen kuuluminen näkyy menettelyssä siinä, että kaikkien osioiden latauksen itseisarvo ensimmäisellä rotatoimattomalla pääkomponentilla on "riittävä". Kommunaliteetin alku- ja loppuarvo ovat hyviä lähtökohtia muuttujan riittävän yhteisen osuuden arvioimiseen. Itse analyysissa näet miten muuttujat löytyvät faktoreilta. Muuttuja joka ei omaa yhteistä muiden kanssa menee usealle faktorille ja/tai jää lopulliselta kommunaliteetiltaan matalaksi. Huonon muuttujan poistamiselle löytyvät nyt perusteet.

Jos haluat laskea koko osiojoukolle alfan, voit kääntää muuttujien pisteityksen pääkomponentin latauksen etumerkin avustuksella. Usein loogisin syin tehty kääntäminen on vailla empiiristä perustetta. Tee käännetyistä osioista uusia muuttujia tiedostoon, joiden nimestä näkee, että pisteitys on käännetty. Suorita faktorianalyysi kuitenkin kääntämättämillä osioilla. Muuttujien merkitysten kääntyminen on helpompi hallita.

2) Suorita faktorianalyysi yleensä paf-menettelyllä. Se vastaa sitä, mitä klassinen eksploratiivinen faktorianalyysi on. Pca-faktorinti tuottaa korkeampia latauksia, mutta myös vain yhdelle muuttujalle rakentuvia faktoreita. ML- faktorointi tuottaa riittävyyden arvioimiseksi khiin neliö -testisuureen. ML tekee kuitenkin voimakkaita oletuksia muuttujien yhteisjakaumista. Menettely on myös herkkä otoksen koon vaihteluille. Eksploratiivisen faktorianalyysin perinne on kuvauksessa, ei tilastollisessa päättelyssä.

3) Tee rotaatio sekä varimax- että promax-rotaatioilla. Promax jatkaa varimaxista yrittäen parantaa sitä. Se luopuu faktoreiden korreloimattomuuden vaatimuksesta. Saat käsityksen kuinka paljon faktorit korreloivat. Tästä saa käsityksen myös tekemällä faktoripisteet summamuuttujina ja korreloimalla ne keskenään.

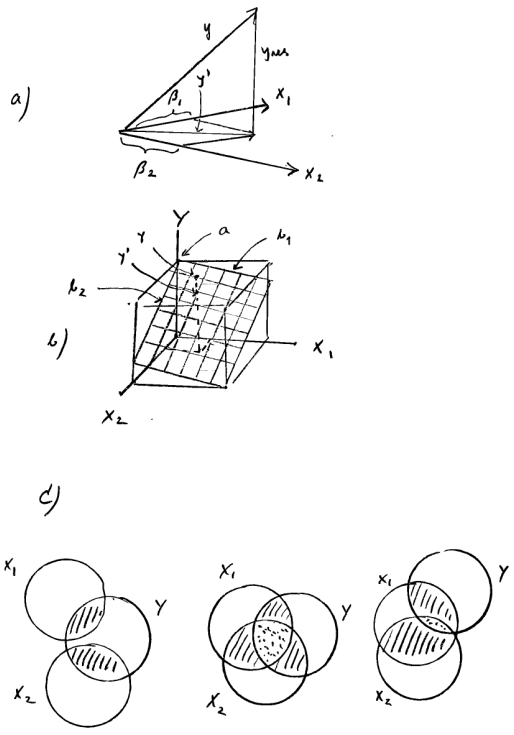
4) Laske reliabiliteetit (alfa-indeksit) faktorianalyysin ohjaamina. Muuttuja voi olla mukana vain yhdessä summassa. Muista muuttujien pisteityksen kääntäminen tarvittaessa. Kun lasket summan reliabiliteetin, saat käsityksen muuttujan (faktorin) reliabiliteetista.

5) Jos käytät regressioestimoituja faktoripisteitä ei osioita tarvitse kääntää saman suuntaisiksi. Menettely ottaa korrelaatiot painokertoimien etumerkeissä huomioon. Faktoripisteiden varianssit (hajonta toiseen) ovat selitysosuuksia (R-

toiseen), jotka saadaan kun muuttujilla estimoidaan faktoreita. Niitä voi käyttää reliaabeliuden arvioinnissa alfojen rinnalla. Muista, että pca:n yhteydessä faktoripisteet voidaan laskea tarkasti (eli hajonta on 1) ja että ne pysyvät täysin korreloimattomina. Promax-rotaatio ja summamuuttujien väliset korrelaatiot kertovat kuinka hyvin korreloimattomuus todella toteutuu.

Promax-rotaatiosta voit valita rakennekertoimet (structure). Ne ovat muuttujan ja faktorin välisiä korrelaatioita. Näet miten vinossa rakenteessa muuttuja voi korreloida korkeasti useankin faktorin kanssa. Vinorotaation yhteydessä faktorikohtaisia selitysosuuksia ei voi laskea yhteen. Pattern-kertoimet antavat saman tiedon kuin lataukset (vain eri esitysmuoto).

Lopuksi regressiomenettelyn visualisointiin:



Kohdassa a) muuttujat ovat vektoreina havaintoyksikköjen avaruudessa. Vektorien pituudet ovat yksi yksikkö. Vektorien väliset kulmien kosinit ovat korrelaatioita. Y-vektorien projektio Xl-X2 -tasolla muodostaa pisteen, jonka yhdensuuntaiset määrittävät betal :n ja beta2:n. Y:n kulma Xl-X2 -tason kanssa on R.
Kohdassa b) havaintoyksiköt ovat pisteparvena (josta merkitty vain yksi kappale) muuttujien koordinaatistossa. Regressioyhtälö on taso, jolla on kulmakerroin X1:n ja X2:n kanssa (b-kertoimet). Taso leikkaa Y:n kohdassa a (vakiokerroin). Regressiotason ja X1-X2 -tason välinen kulma on multippelikorrelaatio. Koordinaatistossa oleva piste on Y, tason määrittämä kohta (estimaatti, Y') ja näiden erotus (Y-Y') on jäännös (residuaa1i).

Regressiotaso kulkee pisteparven läpi pienimmän virheen neliösummavaihtelun periaatteella. Pisteparven hahmottamista auttaa asian kuvittelu kuutiona, jossa yksittäinen piste on X1-X2-Y -koordinaatistossa määritetty.

Kohdassa c) on kuvattu kolme tilannetta. Ensimmäisessä X1 ja X2 ovat korreloimattomia prediktoreita. Korrelaatiot ovat suoraan beta-kertoimia ja yhteiskorrelaation neliö on erillisten korrelaatioiden neliöiden summa- Kohdassa kaksi on tilanne, jossa myös selittävät muuttujat korreloivat. Kohdassa kolme on erityinen ns. suppressiotilanne. X1 korreloi X2:n kanssa mutta korrelaatio Y:hyn on nolla. Tällaisessa tilanteessa X2:n omaosuus ja beta:n etumerkki ovat usein ongelmallisia. Vaikka X2:lla ei ole suoraa tekemistä Y:n kanssa, muodostaa se yhdessä X1:n kanssa regressioyhtälön, jossa se saa merkittävän roolin (suora korrelaatio pieni, omaosuus suuri, beta:n ja suoran r:n etumerkki usein poikkeavat toisistaan). Tulkintakin usein on vaikeaa.

Kohdan c) kuvaus ei ole havainnollistuksena ongelmaton. Tällä tavalla on mahdoton kuvata tilannetta, jossa X1 ja X2 korreloivat negatiivisesti mutta X1 kriteeriin positiivisesti ja X2 negatiivisesti. Tilanne, jossa suoran korrelaation ja betan etumerkki on eri, on käyttäjän kannalta hämmentävä ja epälooginen.

Täydennykseksi faktoripisteihin on syytä mainita, että Lauri Tarkkonen (1987) ja Kimmo Vehkalahti (2000) ovat kehitelleet yhdistelmämuuttujan luotettavuuden arviointia merkittävällä tavalla. Cronbachin alfan käyttämisessä on merkittäviä puutteita. Tätä kirjoitettaessa sovellus on saatavilla Survo-ohjelmassa (www.survo.fi). Ohjelma on yliopistolaiselle erittäin edullinen.

II Tilastollinen päätöksenteko

Tähän saakka olemme käsitelleet menetelmiä, joiden avulla aineisto pyritään saattamaan helpommin käsitettävään muotoon. Juuri tätä vartenhan lasketaan keskiarvoja ja korrelaatioita, tehdään ristiintaulukointeja ja faktorianalyysejä jne. Alkuperäinen materiaali on yleensä liian laaja ja sekava, jotta siitä saisi hyvän yleiskuvan. Olemme siis käsitelleet tilastollista kuvausta (engl. descriptive statistics).

Tutkija joutuu kuitenkin usein tilanteisiin, joissa ei riitä, että hän kykenee kertomaan ja kuvaamaan jonkin muuttujan arvon jakautumisesta tai muuttujien välisten yhteyksien olemassa olosta ulottuen vain siihen joukkoon havaintoyksiköitä, joita hän on tutkimassa. Kaikessa tutkimuksessa (tapaustutkimuksessakin, engl. case-study) on pienenä intressinä mahdollisesti yleisempää mielenkiintoakin koskevan tiedon saavuttaminen.

Kvantitatiivisessa tutkimuksessa voidaan kysyä, kuinka tarkka saatu kuvaus on suhteessa siihen, että tutkittava ryhmä olisikin ollut huomattavasti suurempi. Luottamusrajojen arvioiminen on tällaista arviointia. Mille alueelle tutkimamme ryhmän perusteella muuttujan todellinen arvo suuremmassa joukossa havaintoyksiköitä.

Tutkija ei yleensä saa käsiinsä tai ei käytännön syistä (raha, aika) pysty/halua tutkia koko sitä joukkoa, joka tosiasiallisesti on mielenkiinnon kohteena. Tällainen kohde on luonteeltaan perusjoukko (engl. population). Perusjoukosta peräisin olevalla otoksella (engl. sample) hän pyrkii päättelemään, mikä on asioiden tila perusjoukossa. Otoksen ja perusjoukon suhde on päättelyn kannalta erittäin tärkeä. Erilaisilla otantatavoilla pyrimme huolehtimaan siitä, että tilanne ei ole harhainen (engl. biased) vaan otos edustaa kunnolla perusjoukkoa. Vaaligalluppia varten kännykän kautta haastatellut eivät edusta perusjoukkoa, koska lankapuhelimen omistajat ja puhelittomat aiheuttavat häiriötä otoksen edustavuuteen. Ja taitaapa yleensä olla aika vaikea tehdä otos siitä perusjoukosta, joka ovat ne äänivaltaiset, jotka myös vaalipäivänä käyttävät äänensä. Ovensuukyselytkin ovat epäedustavia, koska postiäänestäjät eivät käy vaalihuoneistoissa.

Vaikeaa on siis edustavuuden saavuttaminen otokselle. Siihen kumminkin pitää pyrkiä. Varmistaa sitä ei voi. Jos voisi, niin otantasattumasta ei tarvitsisi puhua.

Käytännössä (esim. pro-gradu -tutkielmissa) yleistettävyydeltään hyviin otantaratkaisuun perustuvat tutkimuskohteiden valinnat ovat harvinaisia. Kasvatustieteilijän otantaratkaisu on usein käytännön sanelema. Mistä saan tutkimusluvan siellä tutkin. Tutkimushenkilöt sijoittuvat monasti alueisiin, kouluihin ja luokkiin. Perusmenetelmillä näitäkään kontekstivaikutuksia ei juuri vai hallita. Raadollinen fakta on, että liki aina meidän on puhuttava näytteestä, jonka olemme sattuneet saamaan käyttöömme, ei todellisesta otoksesta.

Voiko tilastollinen päätöksenteko, estimointi, luottamusrajojen arviointi ja yleistäminen (kuviteltuun) perusjoukkoon olla relevanttia tällaisessa yhteydessä? Sokeaa tilastomenetelmien soveltamista ja uskoa tunnuslukuihin tulee siis välttää ja käyttää saatua tietoa varovaisesti, tietoisena yleistämisen
sudenkuopista. Yksi sellainen on postikyselyn kato. Katotarkastelu tulisi
aina suorittaa. Käytännössä se usein on vaikeaa, koska perusjoukosta ei juuri tiedetä mitään. Yrittää kuitenkin pitäisi. Tutkittavaan otokseen valinta ja valikoituminen on tärkeä pohdintakohde.

Oletetaan aluksi vaikka tilanne, jossa tutkija on hankkinut käyttöönsä kymmenen helsinkiläistä peruskoulun päättöluokkaa tarkoituksenaan selvittää, millainen on sanallisten laskutehtävien ratkaisemisen taso. Hänellä on siis noin 250 oppilasta ja heidän testituloksensa. Näiden tietojen avulla tutkija laskee keskiarvon ja sille luottamusrajat. Tuloksen hän väittää kuvaavan helsinkiläisten peruskoulua päättävien oppilaiden suoritustasoa. Suoraan ei näinkään suuresta näytteestä voi väittää, että saatu keskiarvo olisi likimain sama kuin perusjoukon arvo. Tietyn kokoiset poikkeamat perusjoukon arvon ja otoksen arvon välillä ovat aivan hyvin mahdollisia. Tällaisten erojen suuruusluokkaa arvioidaan otantajakaumien avulla. Intuitiivisesti tuntuu selvältä, että muutamalle oppilaalle perustuva keskiarvo on ailahtelevampi arvio perusjoukosta kuin useisiin kymmeniin tai jopa satoihin oppilaisiin perustuva keskiarvo. Näin on. Otantasattuman osuus pienenee otoksen koon kasvaessa. Monissa tapauksissa pieneneminen noudattaa suhdetta l/neliöjuuri(N).

Otantajakauma on se asia, jonka avulla voimme arvioida kuinka suuria poikkeamia on odotettavissa otoksen ja perusjoukon (tuntemattoman) arvon välillä kulloisessakin tilanteessa.

Kuvatussa tilanteessa voimme laskea saadulle keskiarvolle esim. 95 %:n luottamusrajat (CI = Confidence Intervall). Se on meidän esityksemme siitä, missä populaation arvo on tietyllä luottamuksella (tai riskillä olla väärässä). Yleistämiseen liittyvänä ongelmana jää elämään se, ovatko luokat edustavia opettajien, opetuksen, koulun varustetason, alueen sosioekonomisen tason yms. suhteen. Vaativissa suurissa evaluaatiotutkimuksissa nuokin seikat otettaisiin tarkastelussa huomioon.

Toinen tutkija on voinut hankkia tietoa, kuinka korkea on korrelaatio lukemisnopeuden ja oppilaan vanhempien välisen koulutuksen määrän välillä. Oletetaan nyt, että molemmat muuttujat ovat kohtuullisen hyvin kvantitatiivisia pseudointervallisia muuttujia ja että yhteyttä voi luonnehtia lineaariseksi. Jos korrelaatiokerroin otoksessa on vaikkapa .20, herää kysymys onko se nyt vain otantaan liittyvä sattuma-asia vai voidaanko väittää, että otoksen takana olevassa perusjoukossakin yhteys poikkeaa nollayhteydestä.

Korrelaatiokertoimenkin yhteydessä voidaan ajatella samanlaista menettelyä. Arvioidaan kuinka paljon vaihtelua on peräkkäisistä otoksista saaduissa korrelaatioissa (korrelaatiokertoimen otantajakauma). Laaditaan tätä tietoa hyväksi käyttäen haarukka eli luottamusrajat sille, missä rajoissa arvelemme tätä otosta vastaavan perusjoukon korrelaation voivan (tietyllä erehtymisen riskillä) olla. Jos nolla kuuluu tuohon luottamusväliin päädymme hyväksymään sen mahdollisena arvona emmekä pidä saamaamme arvoa .20 tilastollisesti merkitsevänä. Tutkija joutuu siis tekemään päätöksen: otoksen arvo joko viittaa yhteyden olemassaoloon perusjoukossa tai sitten ei. Korrelaatiokerroin on tilastollisesti merkitsevä tai sitten ei-merkitsevä - toisella ilmauksella sanottuna.

Kysymys edellisessä kuvatuissa tutkimustilanteissa on tilastollinen merkitsevyystestaus (tapahtui se sitten suorana merkitsevyyden testaamisena tai epäsuorasti luottamusrajojen kautta) tai hieman laajemmin, tilastollisesta estimoinnista ja päättelystä (engl. inferential statistics).

Kuvaavaa ja päättelevää puolta ei voi toisistaan jyrkästi erottaa. Tässä kirjassa ero on tehty vain pedagogisista syistä. Käytännössä kuvaus yleensä sisältää myös merkitsevyyttä koskevia päätelmiä. Samoin merkitsevyyteen pitää yleensä kyetä liittämään kuvaus yhteyden selvyydestä. Tällä selvyydellä on hieman vanhahtava nimi "vaikutuksen koko" (engl. effect size, joka viittaa kokeelliseen tutkimukseen). Yhteyden selvyyttä koskevia tunnuslukuja edellisissä luvuissa olivat esim. multippelikorrelaation neliö ja omaosuus (eli osakorrelaation neliö). Vastaavalla tavalla voidaan asia ilmaista keskiarvoja vertailtaessa (esim. varianssianalyysin yhteydessä eta-toiseen) ja muissakin tilastollisissa merkitsevyyden testaamisen tilanteissa.

Otantajakauma on keskeinen käsite tilastollisessa päättelyssä. Sen ymmärtäminen on tärkeä askel asian hallinnassa.

1. Normaalijakauma

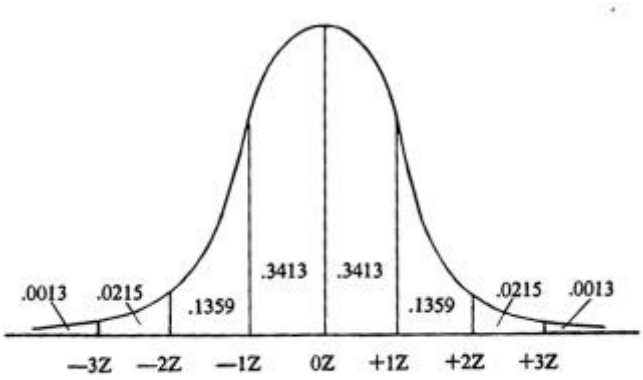
Gaussin kello-käyrästä on puhuttu paljon. Jakautuuko älykkyys normaalisti ? Oletetaan osiotasoksi kymmeniä pieniä osatehtäviä, jotka pisteitetään 2-portaisella skaalalla (oikein/väärin). Mittarin osioiden summapistemäärä kuvaa mittarissa (esim. nykyinen osa kauppakorkeakoulujen yhteistä kirjavalintaa) menestymistä. Summapistemäärä, joka perustuu osioihin joiden korrelaatiot toisiinsa eivät ole kovin korkeita, tulee muodostamaan normaalijakaumaa lähestyvän jakauman sitä tarkemmin mitä enemmän osioita on. Tämän lainomaisuuden esitti jo aikoja sitten eräs suomalainen matemaatikko (J. W. Lindeberg 1920-luvulla). Älykkyys ei ehkä jakaudu normaalisti, mutta älykkyyttä mitataan sillä tavalla, että menettely tuottaa lopputuloksenaan normaalijakauman. Mutta tuskin kiista älykkyyden jakautumisesta loppuu tähän tosiasiaan.

Otetaan toinen esimerkki, mutta otantaan paremmin liittyvä. Oletetaan, että uimataitoisiksi itseään arvioivia on suomalaisissa 70% (populaation, perusjoukon arvo). Otamme ison määrän otoksia joiden kunkin koko on vaikkapa N=140 ja laskemme sen perusteella joukon perusjoukkoa edustavia %- lukuja. On aika ymmärrettävää, että osa luvuista on lähellä 70 %, mutta otantasattumasta johtuen (eli keitä on yksinkertaiseen satunnaiseen otokseen sattunut joutumaan) yksittäiset %-luvut vaihtelevat jonkin verran tuon 70:n %:n ympärillä. Vaikka alku-

peräinen jakauma on 30/70% (vain kaksiarvoinen, kyllä tai ei yksilön kohdalla) on otoksista saatujen %-lukujen jakauma kumpumainen, yksihuippuinen, symmetrinen, normaalia muistuttava jakauma. Monista toisistaan riippumattomista yksiköistä laskettu yhdistelmä (keskiarvo) jakautuu sekin normaalisti olkoon perusjoukon jakauma vaikka dikotominen ja vino.

Molemmat edellä kuvatut jakaumat ovat siis likipitäen normaalijakaumia: yksihuippuisia, symmetrisiä (ei suuresti vinoja, skewness) ja huipukkuudeltaan (kurtosis) Gaussin käyrän tapaisia. Todennäköisyyksien arviointiin käytettävä normaalijakauma on puolestaan matemaattinen Gaussin käyrä. Teoreettisen käyrän ja näiden käytännön sovellusten välinen yhteen käyvyys on kuitenkin riittävän hyvä teoreettisen käyrän ominaisuuksien soveltamiseksi todennäköisyyksien tai riskien arvioimisessa käytännön tutkimustilanteissa.

Normaalijakaumaa ei sinällään ole varattu kumpaankaan arviointiin. Sen ominta aluetta on kuitenkin jälkimmäinen eli otantajakauma. Tutustumme nyt lähemmin normaalijakaumaan. Sen pinta-aloja voidaan rinnastaa todennäköisyyteen ylittää joku arvo tai olla jollakin arvovälillä. Lähes koko pinta-ala sisältyy kolmen hajonnan mitan päähän keskiarvon ala- ja yläpuolella. Standardipisteinä esitetty jakauma ei ole sidottu mihinkään tiettyyn keskiarvoon ja hajontaan vaan on yleinen esitys teoreettisesta normaalijakaumasta. Teoreettisesta jakaumasta pinta-aloja voidaan arvioida tarkkuudella, jossa suhdeluvussa on neljäkin desimaalia. Samoin tiettyä p-arvoa vastaava Z:n arvo voidaan ilmaista samalla tarkkuudella. Tilastolliseen riskin ilmaisemiseen ovat traditionaalisesti kuuluneet kynnystodennäköisyydet .05, .01 ja .001.



Jakaumaan lasketut arvot ovat todennäköisyyksiä, p-arvoja, joista saa prosentteja kertomalla ne sadalla. Niinpä esim. todennäköisyys olla alle -3:n Z-pisteen on .0013 tai 0.13 %. Keskiarvon ja +1:n standardipoikkeaman väliin jääviä arvoja on 34.13 % kaikista, mikä on sama asia kuin todeta, että tälle alueelle osumisen todennäköisyys sattumalta on .3413. Jakauman pinta-alojen osuudet koko jakaumasta ovat siis samalla todennäköisyyksiä osua näille alueille.

Aluksi esitetyissä kahdessa sovellustilanteessa jakauman käytössä on kuitenkin hyvin selvä periaatteellinen ero. Edellinen on kuvaus siitä, miten otoksen YKSITTÄISET pistemäärät jakautuvat eli otoksen jakauma. Jälkimmäinen on otoksista tuotettujen KESKILUKUJEN jakauma eli otantajakauma. Kummankin ominaisuuksia voidaan tarkastella normaalijakauman avulla. Edellisessä keskiarvo ja hajonta kertovat jakaumasta paljon. Tieto huipukkuudesta ja vinoudesta täydentää kuvausta. Yksittäinen muuttujan arvo voidaan muuttaa Z-pistemääräksi, jolloin sen etäisyys keskiarvosta (suuntaan tai toiseen) tulee arvioiduksi hajonnan määränä. Suhteellinen sijainti tarkentuu normaalijakauman ominaisuuksien avulla.

Jälkimmäisessä tilanteessa kaikkien %-lukujen (jotka ovat kukin jo keskiarvo) keskiarvo (kun otoksia on kymmeniä tai satoja) tulee hyvin lähelle perusjoukon arvoa. Jakauman hajontaa taas nimitetään prosenttiluvun keskivirheeksi. Tuolla hajonnalla siis %-luvut vaihtelevat pidemmän päälle kun yksittäisen otoksen N=140.

Jos otamme perusjoukosta (populaatiosta), jonka keskiarvo on 100 ja hajonta (standardipoikkeama) 15 otoksen (N=10) ja saamme sen keskiarvoksi 84 pistettä, olemme tekemisissä otantajakauman kanssa. Jos esitämme kysymyksen: kuinka usein tällainen poikkeaman syntyy sattumalta, otantasattuman seurauksena, kysymme erotuksen 84-100 yleisyyttä tai harvinaisuutta, kun N=10. Nyt emme punnitsekaan sitä hajonnalla 15, vaan otantajakauman hajonnalla (keskivirheellä), joka olisi tässä tilanteessa noin 4.7 (myöhemmin ymmärrät, mistä tämä arvo tulee). Z-piste onkin nyt -3.4 eli aika kaukana "keskiarvosta" (otantajakauman hajonnan mittakaavassa). Tiedämme, että tuo tai tuotakin suuremmat erotukset otoksen keskiarvon ja perusjoukon arvon välillä ovat sattumasyistä (otanta) siis hyvin harvinaisia. Tilastollisessa päätöksenteossa hyl-

käisimmekin sen ajatuksen että kyseinen otos on tästä perusjoukosta ja sanoisimme että erotus on tilastollisesti merkitsevä. Otamme kyllä pienen riskin olla väärässä. Ilmaistaan sama asia toisella tavalla. Otamme yhden 150 henkilön umpimähkäisen otoksen. Saamme %-luvun arvoksi 84. Laskemme tälle %-luvulle luottamusrajat. Perusjoukon oletettu arvo 70 % (jos emme tietäisi sitä) ei sisälly luottamusrajoihin. Hylkäisimme sen hypoteettisena arvona. Jos se olisi tosi, olisi tällainen otoksen arvo kovin harvinainen.

Yleensä meillä ei ole tietoa perusjoukon arvosta. Otoksen arvoa tarjoaa meille mahdollisuuden punnita olisiko joku hypoteettinen arvo mahdollinen perusjoukon arvoksi. Arvot luottamusrajojen sisällä ovat mahdollisia. Arvot sen ulkopuolella "mahdottomia" (periaatteessa mahdollisia, mutta niin harvinaisia että hylkäämme hypoteesin hyväksymisen ajatuksen, ja otamme riskin olla väärässä, hylätä väärin perustein, alfa-tyypin erhe, hylkäämiserhe).

Kun asia on hallussa, pystyt (myöhemmin) vastaamaan seuraavaan. Mikä on todennäköisyys että yllä mainitusta populaatiosta (KA = 100, HAJ = 15) olevan otoksen keskiarvo on 105 tai suurempi, kun otoksen koko N=100? Piirrä otantajakauma ja sijoita havaittu arvo siihen. Osoita alue normaalijakaumasta, joka on vastaus kysymykseen. Palaa tähän, kun asia tulee esille keskiarvon keskivirheen kohdalla.

2. Binomijakauma

Tässä yhteydessä on syytä mainita myös binomijakauma. Se esiintyy jakaumana joskus, joten jotain siitä on hyvä tietää. Binomijakauman elementteinä (toistoina, alkioina) ovat yksiköt, joilla on kaksi vaihtoehtoa esim. kruuna/klaava tai oikein/väärin. Edellisessä on 50/50 tilanne toisessa ei välttämättä. Jos vaikkapa kuvittelisimme tentin, jossa on sata tehtävää. Ne pisteitetään vain oikein/väärin. Oletamme, että kaikki vastaajat ovat pelkästään arvaamalla liikkeellä. Vastauksista syntyy binomijakauma. Jos vaihtoehtoja on vain kaksi syntyy n. 50 pisteen tuntumassa oleva keskiarvo. Sinne sijoittuu jakauman huippu. Mitä kauemmaksi tästä poiketaan, sitä harvinaisemmiksi pelkästään arvaamalla saadut pisteet ovat. On vielä varsin mahdollista saada arvaamalla 55 pistettä, mutta 90 tai korkeamman tai 10 tai matalamman pistemäärän saavuttaminen on jo liki mahdotonta.

Binomijakauma on teoreettisella 50/50 kohdalla varsin lähellä normaalijakaumaa kun alkioita on runsaasti. Se on kuitenkin aina epäjatkuva (diskreetti) ja sen avulla voidaan vastata sellaiseenkin kysymykseen, kuinka todennäköistä on saada täsmälleen pistemäärä 53. Mitä suurempi on vaihtoehtoisten yksiköiden (toisto) määrä, sitä lähemmäksi normaalijakaumaa päästään. Asiaan vaikuttaa myös arvaamisen teoreettinen todennäköisyys. Jos vain yksi vaihtoehto viidestä (esim. monivalintatehtävä) olisi oikein, pakkautuvat pisteet (100 tehtävän tilanteessa) kahdenkymmenen tienoille. Jakauma jäisi positiivisesti vinoksi. Tuo normaalijakaumaa muistuttava piirre on kuitenkin niin vahva, että binomijakaumaa ei itseään tarvitse laskea. Sen keskiarvolla ja hajonnalla pärjää aika hyvin portaittaisuudesta tai vinoudesta huolimatta.

Tarvitaan siis keskiarvo ja hajonta, jotta binomijakauma olisi kohtuullisesti määritetty. Ne saadaan seuraavasti:

$$\mu_b = n * p$$

$$\sigma_b = \sqrt{n * p * q}$$

Kreikkalaiset kirjaimet myy ja sigma kertovat, että on kyse jakauman teoreettisista arvoista, äärettömäksi kuviteltujen toistomäärien kohdalla toteutuvista arvoista, parametreistä erotukseksi otosten arvoista tai otosten arvojen perusteella tehdyistä arvioista, estimaateista. n on toistojen määrä ja p on suotuisan lopputuloksen todennäköisyys yksittäisen toiston kohdalla, q on sen vastatapahtuma 1-p.

Yksittäisen dikotomisen muuttujan (0/1) otoksesta laskettu arvo on empiirisesti saatu p. Se on samalla osion keskiarvo. Se kertoo kuinka moni tapaus kaikista (suhdelukuna) edustaa ykköstä. Dikotomisen muuttujan hajonta on sqrt(pq) ja varianssi pq.

Voimme nyt määritellä edellä mainitun todennäköisyyden saada arvaamalla 55 pistettä tai enemmän, kun tehtäviä on sata. Oletetaan että tehtävät ovat vaihtoehtotehtäviä, joiden p ja q ovat .5. Jotta voisimme käyttää normaalijakauman pinta-alaa (ja sen taulukkoa), meidän on muutettava kyseinen pistemäärä Z-pisteeksi. Edellä esitetyn kaavan mukaan keskiarvo np=100*.5=50. Se tuntuu sopivan aivan arkiajatteluunkin. Toisin sanoen, todennäköisintä pidemmän päälle on saada puolet tehtävistä oikein. Hajonta on

sqrt(npq)=sqrt(100*.5*.5)=5. Voimme siis todeta, että 55 pistettä on yhden hajonnan mitan päässä keskiarvosta. Se vastaa Z-pistettä +1.0. Todennäköisyys yltää tähän tai vielä korkeammalle jakaumassa on yhden hajonnan mitan ylittävä pinta-alan osa normaalijakaumaa. Taulukosta tai edellä esitetyistä kuvioista voimme todeta, että tämä todennäköisyys on 15.87 %. Noin 16 % riittänee vastauksen tarkkuudeksi. Tässä ei oteta huomioon binomijakauman portaittaisuutta. 18% näyttäisi olevan binomijakaumasta saatu arvo. Hyvä olisi huomioida että luokka 55 alkaa sen alarajasta 54.5 jatkuvassa jakaumassa. Z-pisteellä 0.9 tulisi tarkempi tulos normaalijakaumastakin.

Binomijakaumaa käytetään siis vaikkapa monivalintatehtävien arvausjakauman arvioimiseen. Otantajakaumana sen käyttö on vähäistä. %- lukujen merkitsevyyden yhteydessä sitä saattaa nähdä käytettävän. Koska se menee hyvin yhteen normaalijakauman kanssa melko pienelläkin toistojen määrällä (toisto on tässä sama kuin osatehtävä) ei sille jää juuri käyttöä. Binomijakauma teoreettisestikin yhtyy normaalijakaumaan, kun toistojen määrä kasvaa äärettömän suureksi. Tilastolliset asiat ovat arviointiasioita. Niiden yhteydessä eksakti laskeminen on toissijaista. Normaalijakauman ominaisuuksia voidaan siis käyttää hyväksi.

Binomijakaumaa otantajakaumana käytettäessä voidaan toistojen määrä rinnastaa otoksen havaintoyksiköiden (joista kullakin arvo 0 tai 1) määrään. Näistä laskettu keskiarvo vaihtelee otoksesta toiseen (jo muutaman kymmenen otoskoolla) varsin lähellä normaalijakaumaa olevalla tavalla. Keskivirhe (otantajakauman hajonta) voidaan arvioida käyttämättä binomijakauman tarkkaa laskentaa (jota ei olekaan tässä kirjassa). Edellinen kaava ei sellaisenaan käy. Dikotomisen muuttujan hajonta on neliöjuuri(p*q). Tämä jaettuna neliöjuuri(N) -termillä antaa jonkinlaisen arvion otantajakauman hajonnasta (keskivirheestä).

Ja myöhempää varten seuraavia asioita. Normaalijakauman avulla tehdään merkitsevyystestaus tutuksi. Kellokäyrää ei myöhemmissä sovelluksissa sitten juurikaan käytetä. t-jakauma korvaa sen, koska otantajakaumalla taipumus olla "leveämpi" kuin normaalijakauma. Normaalijakauma on siis erityistapaus t-jakaumaa, kun otoskoko (teknisemmin: vapausasteet, degrees of freedom, df) kasvaa suureksi. t- jakauman voidaan taas sanoa olevan erityistapaus F-jakaumaa. Khii-toiseen jakauma tulee tutuksi erityisesti ristiintaulukoiden yhteydessä. Se yhtyy vapausasteella df=l normaalijakaumaan (khii-toiseen = z-toi-

seen). Khiin neliön käytön yhteydessä tullaan havaitsemaan, että vapausasteet liittyvät käsitteenä sen yhteydessä ristiintaulukon solukon (esim. 2 * 3 taulukossa on 6 solua) kokoon. Löytyy koko joukko muitakin sovelluksia.

Käytännössä tietokonetulostuksissa tulee kohtaamaan siis yleensä F-arvoja tai khii-toiseen arvoja. Niiden ilmoittamiseen ja käyttöön liittyy aina tietty tai tietyt vapausasteet.

Binomijakaumasta voisi otantana tehdä jonkinlaisen (aika keinotekoisen) esimerkin. Tutkimusaineisto on 80 yksilapsista perhettä. Lapsen biologinen sukupuoli koodataan l=tyttö, 2=poika. Arvioi sukupuoli-muuttujan otantajakauma. Kuinka mahdolliseksi arvioit keskiarvon 1.62 (kun teoreettinen odotus on n. 1.50? "Noin" siksi että muistaakseni poikia syntyy lievästi odotusta enemmän (mutta se näkynee vasta erittäin suurissa luvuissa).

Koodatun muuttujan desimaaliosalla on mukava ominaisuus. Sen desimaaliosa kertoo suoraan korkeamman koodin osuuden koko joukosta (eli tässä 62 % on poikia). Sukupuoli-muuttujaa pystyy siis käyttämään tulomomenttikertoimenakin (se on ns. piste-biseriaalinen korrelaatio). Kolmeluokkaista kvalitatiivista muuttujaa sen sijaan ei sellaisenaan voi käyttää keskiarvon, hajonnan ja korrelaation laskuissa.

3. Otantajakauma

Yleensä tutkija ei ole kiinnostunut yhteen yksilöön liittyvistä todennäköisyyksistä, vaan siitä, kuinka lähellä hänen otoksensa tunnusluvut, esim. keskiarvo, ovat populaation todellisia arvoja. Kyseessä on siis useiden yksilöiden muodostaman joukon asema populaatioon nähden. Tällöin muuttuvat todennäköisyydet oleellisesti äskeiseen verrattuna. Voimme tarkastella tätä ensin keskiarvojen todennäköisyyksien kannalta.

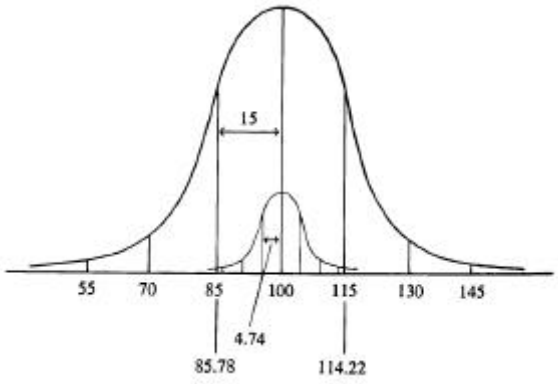
Aiemmin esitetystä normaaliksi oletetusta älykkyyden jakaumasta voimme todeta mm. että yhden sattumanvaraisesti valitun yksilön todennäköisyys osua yli + 1 :n standardipoikkeaman olevalle alueelle on .159, siis hieman alle 16 %. Minkälainen olisi suuremman joukon, esim. kymmenen henkilön otoksen, keskiarvon osumisen todennäköisyys samalle alueelle? Olisiko se sama vai suurempi vai pienempi? Jotta keskiarvo voisi olla tätä suuruusluokkaa, täytyisi kaikkien tai melkein kaikkien yksityisten pistearvojen olla samaa luokkaa. Tämähän on selvästikin varsin epätodennäköistä. Olisi varsin kummallista, että kymmenen sattumanvaraisesti valittua henkilöä olisivat kaikki huippuälykkäitä. Tarkemmin ajatellen huomaa, että tämä on sitä epätodennäköisempää, mitä suuremmasta joukosta on kysymys. Täytyy selvästikin olla niin, että mitä suurempi otos otetaan, sitä varmempaa on, että sen keskiarvo on lähellä populaation keskiarvoa.

Tähän yleisesti muotoiltuun näkemykseen on olemassa selvä sääntö, jonka käsittelemiseksi tarvitsemme otantajakauman (engl. sampling distribution) käsitettä. Kuvitellaan, että ottaisimme samasta populaatiosta yhä uudelleen, periaatteessa äärettömän monta, kymmenen henkilön otosta, jolle kullekin laskettaisiin keskiarvo. Näissä keskiarvoissa esiintyisi hieman vaihtelua, välillä ne sattuisivat populaation keskiarvon ala- ja välillä yläpuolelle. Joskus ne osuisivat jokseenkin samaan populaation arvon kanssa. Lopulta huomaisimme, että nämä keskiarvot olisivat muodostuneet populaation keskiarvon ympärille oman jakaumansa, joka olisi normaali mutta selvästi kapeampi kuin yksityisten pistemäärien jakauma populaatiossa. Tämä kapeushan johtuu siitä, että usean henkilön otoksen on huomattavasti epätodennäköisempää olla kaukana populaation keskiarvosta kuin yhden yksilön. Tämä keskiarvojen muodostama kuviteltu jakauma on keskiarvon otantajakauma.

Vaikka käytännössä otantajakauma tunnetaan vain suunnilleen, voidaan se teoriassa (oletuksin) tarkoin määritellä . Keskiarvon otantajakauman keski- arvo sijaitsee samassa kohdassa kuin populaation keskiarvokin. Tämähän on luonnollinen seuraus siitä, että on yhtä todennäköistä saada todellista arvoa suurempia kuin sitä pienempiäkin arvoja; otosten keskiarvot siis sijaitsevat symmetrisesti populaation arvon molemmilla puolin. Otantajakauman hajonta eli keskiarvon keskivirhe (engl. standard error of mean) saadaan seuraavasti:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Omassa älykkyyden jakautumista koskevassa esimerkissämme saamme täten kymmenen henkilön otosten keskivirheeksi 15/sqrt(10)=4.74. Kymmenen henkilön otosten keskiarvojen jakauma on siis leveydeltään vajaa kolmasosa yksityisten pistearvojen jakaumasta. Voimme havainnollistaa tätä piirtämällä molemmat samalle asteikolle:



Voidaan ajatella, että tutkimuksessa saatu kymmenen henkilön otoksen keskiarvo on eräs niistä kuvitelluista keskiarvoista, jotka muodostavat otantajakauman. Nyt me teemmekin todennäköisyyksiä koskevat päätelmämme otantajakauman, emmekä koko populaation jakauman puitteissa. Jos siis on keskiarvosta kysymys, tutkitaan keskiarvojen jakauma, eikä yksityisten pisteiden. Tämähän on luotettavuuden kannalta vain ilahduttavaa: sehän merkitsee, että epävarmuus on pienentynyt. Esimerkiksi yli 99 % kaikista kymmenen henkilön

keskiarvoista sijaitsee alueella, joka on -3:n ja +3:n otantajakauman standardipoikkeaman välissä. Kolme standardipoikkeamaahan on 3*4.74 = 14.22 pistettä. Saamme tulokseksi, että otamme vajaan prosentin riskin olla väärässä, jos oletamme sattumanvaraisesti valitun kymmenen henkilön otoksen keskiarvon olevan välillä 85.78 … 114.22 pistettä.

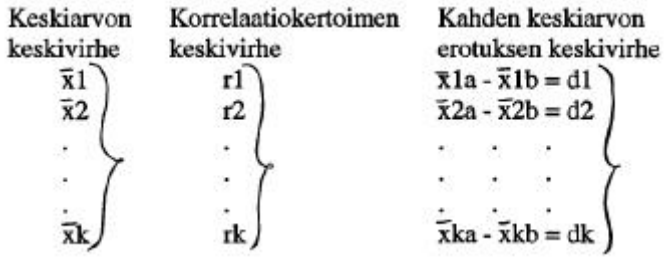
Epävarmuuden pienenemisen huomaa, kun muistaa, että sama riski yksityisten pistearvojen kohdalla johti väliin 55 … 145 pistettä. Jos meillä olisikin ollut sadan henkilön otos, olisi keskivirhe ollut enää 15/sqrt(100)=1.5 pistettä. Runsaat 99 % kaikista tapauksista olisi tällöin sijainnut 100 plus/miinus 3*1.5 =välillä 95.5 … 104.5 pistettä.

Tähän jatkoksi on todettava se, että otantajakauman arvioimissa on tässä oletettu, että tiedämme perusjoukon keskiarvon ja hajonnan. Yleensä sellaisia tietoja ei ole käytettävissä. Tärkeä tieto, perusjoukon hajonta, korvataan otoksesta saatavalla hajonnalla. Voidaan osoittaa, että otoksesta laskettu hajonta on perusjoukon hajonnan harhaton estimaatti. Yksittäisen otoksen kohdalla estimaatti saattaa olla puoleen tai toiseen virheellinen. Virhe ei ole systemaattinen. Hajonnallakin on siis otantajakaumansa (keskivirheensä). Niinpä peräkkäisistä otoksista lasketut keskiarvon keskivirheet vaihtelevat hiukan, sitä enemmin mitä pienemmästä otoksesta on kyse.

Perusjoukon keskiarvoa ei meidän tarvitse tietää. Kun olemme laskeneet otoksen keskiarvon ja keskihajonnan, voimme kääntää päättelyn toiseen suuntaan. Kysymys kuuluu: Mitkä ovat ne rajat, joiden väliset keskiarvot voisivat olla tämän otoksen takana olevan perusjoukon (mahdollisia) arvoja? Tämä on jo selkeä sovellustilanne edellisestä teoriasta. Otoksen perusteella voimme laatia saamamme keskiarvon perusteella luottamusrajat (engl. confidence intervall, CI), joiden välillä otostamme vastaavan perusjoukon arvon uskotaan sijaitsevan.

Samasta perusjoukosta otetuista peräkkäisistä, samankokoisista otoksista saamme kustakin omat luottamusrajat perusjoukon arvolle. Luottamusrajojen keskipiste vaihtelee (koska kukin otos tuottaa samantapaisen, mutta ei täsmälleen samaa keskiarvoa) ja luottamusvälit ovat erilaisia leveydeltään (koska leveyteen vaikuttava hajonta ja sen pohjalta arvioitava keskivirhe vaihtelevat otoksesta toiseen).

Keskivirheen käytännön ajatus kolmessa sovellustilanteessa:



Näiden kolmen tunnusluvun otoksesta toiseen vaihtelevien arvojen hajonta on keskivirhe. Muillakin tunnusluvuilla on keskivirheensä.

4. Luottamusvälin arvioiminen

a) keskiarvo

Äskeisessä selostuksessa olemme puhuneet siitä, miten otantajakauma tarkasti ottaen suhtautuu tunnettuun populaation keskiarvoon ja hajontaan kunkin suuruisissa otoksissa. On kuitenkin helppo todellista tutkimustilannetta ajatellessaan havaita, ettei edellä kuvattu menettely sovi käytännön ohjeeksi: tutkijallahan on käytettävissään vain otoksensa, populaation arvot ovat tuntemattomia. Juuri niiden selville saamiseksi otos on hankittu. Otantajakauma on otosten keskiarvojen jakauma, joka perustuu tunnetuiksi kuviteltujen populaation arvojen (keskiarvo ja hajonta) muodostamalle pohjalle.

Meidän on nyt mietittävä, mitä voidaan tehdä, kun pelkästään otoksen arvot tunnetaan. Esimerkkinä voidaan pitää aiemmin esitettyä 250 peruskoululaisen otosta, jolle tehtiin lukemisnopeutta mittaava testi. Oletetaan, että saamme otoksen keskiarvoksi 50 ja keskihajonnaksi 8 pistettä. Kuinka "tarkka" on saatu keskiarvo, miten kaukana voi populaation keskiarvo olla siitä? Nyt on lähtö-

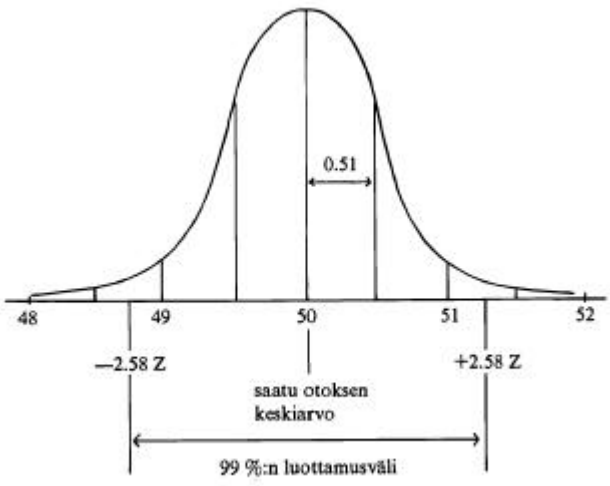
kohtana otoksen keskiarvo ja arvioitavana se alue, missä populaation arvo voisi olla.

Me siis haemme jakaumaa, jonka alueella populaation keskiarvo kohtuullisella uskottavuudella sijaitsee. Menemättä sen tarkemmin asian johtamiseen, voimme todeta, että edellä mainittu jakauma on aivan samanlainen kuin keskiarvon otantajakaumakin. Otantajakauma on populaation keskiarvon kohdalla (joka tunnetaan tai oletetaan tunnetuksi), mahdollisten populaation keskiarvojen jakauma on otoksen keskiarvon ympärillä. Etäisyysmitta kummassakin on sama.

Otantajakauman hajonta (keskivirhe) laskettiin tunnetuksi oletetun populaation standardipoikkeaman avulla. Käytännössä joudumme estimoimaan sen otoksen hajonnan avulla:

$$S_{\bar{X}} = \frac{s}{\sqrt{N}}$$

Voimme nyt esittää kuvan tilanteesta, jossa mainitun 250 henkilön otoksen keskiarvoksi tuli 50 ja hajonta 8. Populaation keskiarvo sijaitsee jakaumassa, jonka hajonta (keskivirhe) on 8/sqrt(250)=0.51 pistettä:

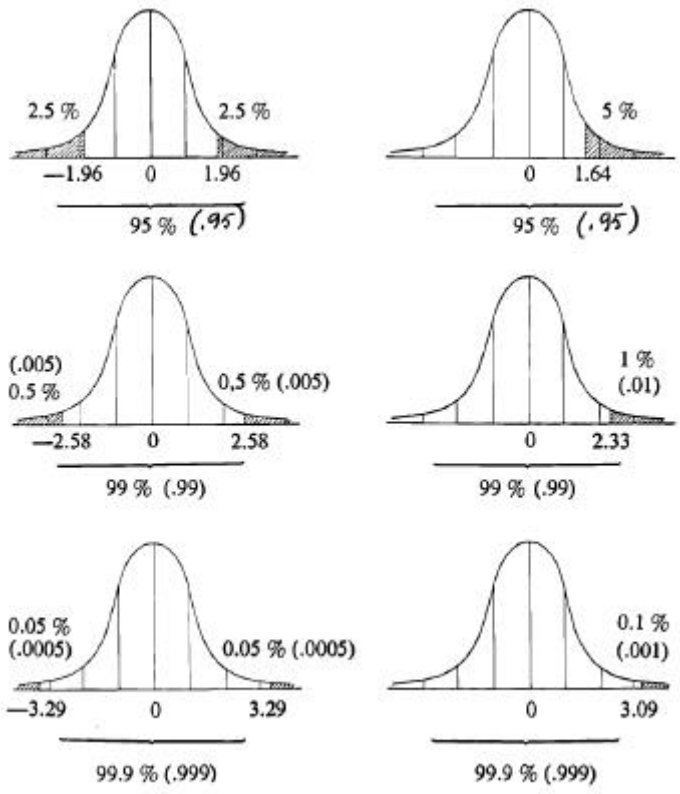


Näin suurella otoksella pääsemme melko suureen tarkkuuteen: populaation keskiarvo sijaitsee 99 %:in todennäköisyydellä noin kolmen pisteen alueella. Tark-

kaan ottaen 99 % tapauksista sijaitsee alueella, joka on 2.58 hajonnan mittaa keskiarvosta kumpaankin suuntaan. Tämä alue on saadun otoksen keskiarvon 99 %:n luottamusväli. Jos siis uskomme populaation keskiarvon sijaitsevan tällä alueella, olemme 99 %:n todennäköisyydellä oikeassa ja otamme 1 %:n riskin olla väärässä. Jos haluamme tietää alueen rajat alkuperäisinä raakapisteinä, kerromme otantajakauman hajonnan (0.51) kriittisellä Z-arvolla (2.58) ja menemme saadun matkan verran keskiarvosta molempiin suuntiin:

$$CI_{.95} = 50 \pm 2.58 * 0.51 = 48.68 \leftrightarrow 51.32$$

Tavallisesti käytettyjä riskitasoja ovat 5 %, 1 % ja 0.1 %, joista siis saamme vastaavasti 95 %:n, 99 %:n ja 99.9 %:n luottamusvälit. Näitä vastaavat kriittiset Z-arvot ovat 1.96, 2.58 ja 3.29, kuten oheisen kuvion vasen puoli osoittaa.



Kuviossa normaalijakaumaan liittyvät todennäköisyydet tavallisimmin käytetyillä riskitasoilla.

Voimme tähän esimerkkinä laskea nämä luottamusvälit vaikkapa kirjan alussa esitetyn primäärimatriisin järkeilytestin keskiarvolle, joka oli 34.97. Kun otoksen numerus oli 30 ja hajonta 1.83, saamme otantajakauman hajonnaksi 0.34. Luottamusvälit tulevat tämän avulla seuraavasti:

$$CI_{.95} = 34.97 \pm 1.96 * 0.34 = 34.30 \leftrightarrow 35.64$$
$$CI_{.99} = 34.97 \pm 2.58 * 0.34 = 34.09 \leftrightarrow 35.85$$
$$CI_{.999} = 34.97 \pm 3.29 * 0.34 = 33.85 \leftrightarrow 36.09$$

Arvoista näkyy selvästi se, kuinka tarvitsemme sitä laajemman alueen, mitä varmempia haluamme olla päätelmän suhteen.

Luottamusvälin voi laskea muillekin tilastollisille tunnusluvuille kuin keskiarvoille. Otamme tässä esiin pari usein esiintyvää tapausta: prosenttiluvun ja korrelaation luottamusvälin laskemisen. Perusperiaate on aivan sama kuin edelläkin: meidän tulee hankkia tieto otantajakauman standardipoikkeamasta (keskivirhe) ja käyttää tätä mittana siitä, kuinka kauas otoksesta saadun arvon kummallekin puolelle meidän on mentävä, jotta haluttu varmuus populaation arvosta saataisiin.

b) Prosenttiluku

Meidän esimerkissämme on yhdeksän sellaista henkilöä, joiden viriketausta on luokiteltu huonoksi. Prosentteina tämä on 30. Kuinka kaukana tai lähellä tämä voisi olla populaation arvoa, miten kauas tästä meidän tulisi mennä, jotta voisimme uskoa vaikkapa 95 %:n varmuudella populaation arvon sijaitsevan annettujen rajojen sisällä? Toisin sanoen, meidän tulee määritellä 30 %:lle luottamusväli 5 %:n riskillä. Prosenttiluvun keskivirhe, siis sen otantajakauman standardipoikkeama, riippuu itse prosenttiluvusta ja numeruksesta seuraavalla tavalla:

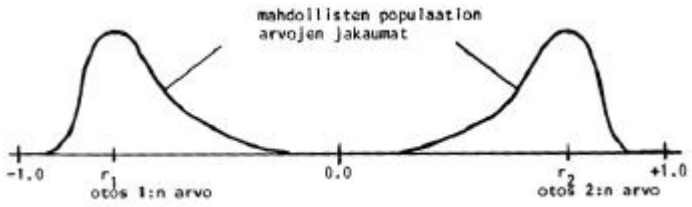
$$S_P = \sqrt{\frac{P * Q}{N}}$$

$$Q = 100 - P$$

Meidän tapauksessamme on siis keskivirhe 8.37. Kun halusimme 95 %:n varmuutta, meidän on mentävä 1.96 standardipoikkeaman verran saadusta otoksen arvosta molempiin suuntiin, siis rajoihin 13.59 … 46.41. Jos uskomme todellisen populaation arvon olevan tällä välillä, otamme 5 %:n riskin, erehtyisimme siis viisi kertaa sadasta. Huomaa, että nyt ovat arvot prosentteja, ts. alempi arvo tarkoittaa, että populaatiossa tuskin on vähemmän kuin vajaat neljätoista prosenttia sellaisia, joilla on huono viriketausta ja, vastaavasti, on epäuskottavaa, että heitä olisi juuri yli 46:n prosentin. Mielipidekyselyissä puolueiden kannatusprosenttien virheen laskeminen tehdään tällä tavalla.

c) Korrelaatio

Korrelaation luottamusvälin laskeminen on hieman mutkikkaampaa kuin edelliset. Tämä johtuu siitä, että korrelaation otantajakauma ja siis myös mahdollisten populaation arvojen jakauma, on vino. Korrelaation arvojen heilahtelujen todennäköisyys kasvaa mentäessä kohti nollaa (perusjoukossa) ja toisaalta, mitä lähempänä ykköstä ollaan, sitä pienempiä ovat todennäköiset virheet. Eri korrelaation arvoja kuvaava asteikko ikään kuin tiivistyy kohti ykköstä. Toisin sanoen, voimme luottaa enemmän korkeaan kuin matalaan korrelaatioon. Näin siis mahdollisten arvojen jakauma on otoksen arvon ympärillä kuten aiemminkin, mutta ulottuu pitemmälle alas kuin ylöspäin, se on vino:



Tämä vaikeus voidaan kiertää muuttamalla korrelaatiot väliaikaisesti normaalisti jakaantuviksi arvoiksi, Fisherin Z-pisteiksi. Toimenpiteet ovat tällöin seuraavat:

- Muunnetaan otoksen perusteella laskettu korrelaatio Fisherin Z-pisteiksi taulukon avulla.
- Lasketaan tälle luottamusväli normaalijakauman mukaan käyttäen keskivirhettä:

$$S_z = \frac{1}{\sqrt{N-3}}$$

- Muunnetaan saadut rajat takaisin korrelaatioiksi taulukkoa käyttäen.

Toisaalta suurella otoskoolla, kun arvioidut perusjoukon korrelaatiot ovat korkeintaan .4 luokkaa itseisarvoltaan, riittää korrelaatiokertoimen keskivirheen arvioksi usein karkea 1/sqrt(N) - ilman mitään muunnoksia.

Voimme esimerkkinä laskea vaikkapa 99 %:n luottamusvälin primäärimatriisissamme olevalle kielten keskiarvon ja viriketaustan väliselle korrelaatiolle .64. Taulukosta saamme korrelaatiota vastaavaksi
Fisherin Z-arvoksi .758. Tämän keskivirhe on 1/sqrt(N-3)=.192. 99 %:n rajojahan vastaavat normaalijakaumassa kohdat, jotka ovat 2.58 standardipoikkeaman päässä jakauman keskeltä. Saamme siis luottamusväliksi Fisherin Z-pisteinä .263 … 1.253. Huomaa, että arvot voivat mennä yli yhden (ne eivät ole korrelaatioita). Kun ne muutetaan taulukon avulla takaisin korrelaatioksi, saadaan .25 ja .85. Näin pienellä tapausmäärällä (30) joudumme siis ottamaan huomioon melko suuren alueen ennen kuin voimme olla kohtuullisen varmoja, että populaation arvo on mukana. Jakauman vinous tulee selvästi näkyviin: alkuperäinen arvo (.64) ei ole saatujen rajojen keskellä.

Sovelluksista ensimmäinen on tavallinen. Keskiarvojen luottamusrajat tulevat jo keskiarvon laskemisessa automaattisesti tulostettua. Standard Error of Mean muodostaa yleisesti käytetyn perustan CI:n laskemiselle. Otantajakauma ei ole

normaalijakauma vaan t-jakauma. Tilastotieteilijä Gossett esitti nimimerkillä "Student" t-jakauman, jota keskiarvon keskivirhe noudattaa. Se on pienellä otoskoolla leveämpi kuin normaalijakauma. Esittämämme esimerkin kohdalla CI olisi pitänyt laskea t-jakaumaa käyttäen vapausasteilla N-1.

Prosenttiluvuilla käytetään joko binomijakaumaa tai muutetaan asia soveltumaan khiin neliön jakaumalle. Myöhemmin huomaamme, että %-lukujen erotuksen merkitsevyydenkin yhteydessä voidaan kaikkein suoraviivaisin tilastollinen testaus tehdä khiin neliöllä.

Korrelaatiokertoimen luottamusrajojen laatiminen on harvinaista. Kahden korrelaation erotuksen merkitsevyyden tarkastelu samoin on harvinainen. Tyypillistä on testata asia siten, että arvioidaan, voisiko otos olla peräisin perusjoukosta, jossa on nollakorrelaatio. Tässä tilanteessa korrelaatiokertoimen otanta palautuu t-jakaumaan. Vapausasteiden lukumäärä on N-2, jossa N on X-Y -parien kokonaismäärä.

Luottamusrajojen arvioinnin periaatteen ymmärtämisen vuoksi asia on esitetty näissä kolmessa tilanteessa. Merkitsevyystestaukseen liittyy yhteyden suuruuden (effektin koko, joka on tekemisissä beta-erheen kanssa), merkitsevyyden (alfa-tyypin erheen riski) ja otoskoon huomioon ottaminen kokonaistilanteeksi. Nämä asiat käsitellään yhdessä kirjan viimeisessä vaiheessa.

5. Kahden tunnusluvun erotuksen merkitsevyys

Hyvin tavallinen tilanne on se, jossa tutkija joutuu päättämään, onko kahden ryhmän välillä todellista eroa. Hän on voinut vaikkapa järjestää kokeen, jossa yritetään vaikuttaa toiseen ryhmään (koeryhmään) ja sen jälkeen tutkitaan, syntyikö ryhmien välille uskottavaa eroa. Tämä ero voi tulla näkyville ryhmien keskiarvoissa, hajonnoissa, joidenkin ominaisuuksien korrelaatiossa ryhmittäin jne. Ryhmät voivat kokeellisen asetelman sijasta olla myös sellaisia luonnostaan olemassa olevia joukkoja edustavia kuten tytöt/pojat, maalla/kaupungissa asuvat jne. Oman esimerkkiaineistomme puitteissa voisimme vaikkapa haluta tietää, onko naisten ja miesten ero järkeilytestissä niin suuri, että voimme hylätä ajatuksen sen syntymisestä sattumalta.

Kahden tunnusluvun erotuksen todellisuutta tutkittaessa on perusperiaate aina sama, vaikka yksityiskohdat vaihtelevatkin. Kuvitellaanpa tilannetta, jossa olisi mahdollisuus poimia samasta populaatiosta hyvin suuri määrä otoksia, jotka olisivat samankokoisia ja muodostuisivat kahdesta ryhmästä, jotka nimitettäisiin aina kussakin otoksessa ryhmiksi 1 ja 2. Tällöin me voisimme laskea kullekin ryhmälle keskiarvon ja kullekin otokselle kahden ryhmän keskiarvojen eron viiva-X1-viivaX2. Vaikka ryhmät onkin poimittu samasta populaatiosta, ei voida odottaa keskiarvojen eron olevan aina nolla tai olevan otosparista toiseen sama. Sattuman takia saisimme välillä vähän suurempia ja välillä pienempiä eroja, välillä olisi ykköseksi nimetyn ryhmän keskiarvo suurempi, välillä kakkosen. Todennäköisin arvo suuressa tapausjoukossa olisi kuitenkin nolla, ts. että eroa ei olisi. Näin meille muodostuisi jakauma, jonka keskikohta olisi nollan (siis: ei eroa) kohdalla, vasen puoli kuvaisi niitä tapauksia, joissa kakkosryhmän keskiarvo olisi ollut suurempi ja siis erotus negatiivinen. Oikealla olisivat vastaavasti tapaukset, joissa ykkösryhmän keskiarvo sattui olemaan suurempi. Otosparien keskiarvojen erotuksista tulisi kumpumainen jakauma keskiarvolla nolla ja hajonta olisi keskiarvojen erotusten keskivirhe. Sen suuruuteen vaikuttaisi paljon, minkä kokoisia otosparit olisivat.

Tämä jakauma, joka osoittaa, miten samasta populaatiosta poimittujen keskiarvoparien erotukset jakaantuvat, on keskiarvojen erotuksen otanta- jakauma. Voimme jälleen ajatella, että tutkimuksessa saatu ero on eräs näistä periaatteessa äärettömän monista eroista. Jakauman keskellä sijaitsevat todennäköiset, helposti sattumaltakin esiintyvät erot; mitä pitemmälle kohti reunoja menemme, sitä epätodennäköisempiä arvot ovat. Jos saamamme ero sijoittuu riittävän kauas reunalle, se tulee niin epätodennäköiseksi, että se ei enää uskottavasti voi olla sattumaa. Tällöin teemme johtopäätöksen, että ero on todellinen. Toisella tavoin sanottuna, katsomme, että ryhmät edustavat tällöin eri populaatioita, erotus on merkitsevä. Jos saamamme ero sijoittuu vaikkapa sen kohdan ulkopuolelle, jota suurempia voi sattumalta saada vain kerran sadasta, sanomme, että ero on tilastollisesti merkitsevä 1 %:n riskitasolla.

Jotta voisimme tietää, mihin kohtaan havaittu ero sijoittuu, on tiedettävä, miten laajalle alueelle erojen teoreettinen jakauma ulottuu; mitaksi on laskettava eron otantajakauman hajonta eli erotuksen keskivirhe. Ero jaetaan tällä keskivirheellä, jolloin saadaan tietää, kuinka monen. standardipoikkeaman päähän "ei eroa tilanteesta " se sijoittuu. On sitten kyse keskiarvojen, hajontojen, prosenttilukujen tms, perusajatus on aina sama. Havaitun arvon ja nollahypoteesin ero jaetaan keskivirheellä. Saatua eroa nimitetään kriittiseksi suhteeksi (CR, critical ratio).

$$\mathrm{CR} = \frac{\text{estimaatin poikkeaminen Ho:n arvosta}}{\text{estimaatin keskivirhe}}$$

Erojen merkitsevyyden testauksessa käytetään yleensä kahta hieman toisistaan poikkeavaa jakaumatyyppiä. Jotkut erot, nimenomaan suurehkojen numerusten ollessa kyseessä, testataan olettaen erotuksen otantajakauman olevan normaali. Tällöin siis jaetaan saatu ero normaaliksi oletetun otantajakauman hajonnalla (eron keskivirhe); tässä tapauksessa puhutaan Z-testistä. Pienehköjen aineistojen kyseessä ollen jakautuvat keskiarvojen erotukset kuitenkin hieman eri tavoin. Tällöin sitä kuvaa parhaiten t-jakauma. Tällaisessa tapauksessa tehdään t-testi: jaetaan ero t-jakaumaan liittyvällä keskivirheellä ja tutkitaan saadun kriittisen suhteen merkitsevyystaso t-jakaumaa kuvaavasta taulukosta.

a) Kahden keskiarvon erotus

Keskiarvojen erojen merkitsevyyttä testataan yleensä t-testillä, joka sopii sekä suurten että pienten aineistojen ollessa kyseessä: äärettömän suurilla aineistoilla t-jakauma on sama kuin normaalijakauma. Tavallisimmassa tapauksessa sopii seuraava kaava t-arvon laskemiseen:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}\sqrt{\frac{N_1 + N_2}{N_1 \cdot N_2}}}$$

Jossa erillisistä hajonnoista muodostettu yhdistetty hajontaestimaatti saadaan:

$$s_{pooled} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}}$$

t-testin (kuten myöhempänä esitetyn varianssianalyysinkin) ajatukseen kuuluu, että ryhmien hajonnat eivät ole huomattavan erisuuret. Jos näin näyttää olevan, on syytä katsoa apua kattavammasta esityksestä. Harvoin hälytysraja kuitenkaan ylittyy (esim. varianssien suhde suurin/pienin yli 4).

Huomaa, että osoittajassa on saatujen keskiarvojen ero, kuten edellä esitetty yleisperiaate edellyttääkin. Nimittäjä on kokonaisuudessaan kaava, jolla saamme eron keskivirheen t-jakaumana. Voimme nyt soveltaa tätä aiemmin esitettyyn kysymykseen siitä, onko naisten ja miesten ero järkeilytestissä niin suuri, että voimme pitää sitä todellisena, ts. onko se tilastollisesti merkitsevä.

Ryhmittäisiä keskiarvoja ja hajontoja ei ole aiemmin laskettu, joten joudumme tekemään sen tässä. Menemättä itse laskuihin, joiden otaksutaan sujuvan, voimme todeta, että miesten keski- arvo on 35.8 ja hajonta 1.8. Naisten vastaavat arvot ovat 34.1 ja 1.4. Ryhmät ovat yhtä suuret, siis N1=N2=15.

Saamme seuraavan laskutoimitusten tuloksena erotuksen 1.7 ja sille keskivirheen 0.61, josta syntyy t-arvo 2.79 vapausasteilla 28. Siinä ollaan siis aika kaukana arvosta, joka olisi odotettavissa sattumatilanteessa.

Ongelmana on nyt enää ottaa selville, kuinka todennäköistä on olla 2.79 hajonnan päässä (tai sitäkin kauempana) t-jakauman keskeltä. Tämän näkee t-jakaumataulukosta, mutta ennen kuin katsomme sen, on huomautettava parista seikasta. Ensinnäkin, koska me olisimme yhtä hyvin voineet joutua jakauman toiseen päähän, jolloin siis naisten tulos olisi ollut parempi, meidän on otettava todennäköisyyksissä huomioon jakauman molemmat päät ja laskettava niiden arvot yhteen (suuntaamaton testaus). Toisin sanoen, jos haluamme ottaa vaikkapa 5 %:n riskin, on otettava huomioon 2.5 % jakauman kummastakin päästä. Jos teemme testauksen 1 %:n riskitasolla, joudumme hakemaan kohdan, jonka ulkopuolelle sijoittuu 0.5 % tapauksista jne. Toinen huomioonotettava seikka on se, että taulukon arvo on haettava kohdasta, joka riippuu ryhmien koosta. Tarkkaan ottaen se riippuu ryhmien niistä havainnoista, jotka ovat riippumattomia, vapaita varioimaan. Tämä on käsitettävä niin, että tietyn tunnusluvun, esim. keskiarvon voi saada hyvin monilla erilaisilla havaintojen arvoilla, mutta kun riittävän moni tunnetaan, ei lopuille enää jää vaihtoehtoja. Keskiarvon tapauksessa on viimeinen havainto täysin määrätty, kun muut tunnetaan. Niinpä yhden ryhmän vapaasti varioivia tapauksia on yhtä vähemmän kuin ryhmän koko, ryhmän vapausasteita (degrees of freedom, df) on N-l. Meidän tapauksessamme, jossa verrataan kahden ryhmän keskiarvoja, on vapausasteita N1+N2-2. On helppo huomata, ettei kakkosen vähentämisellä ole juuri mitään käytännön eroa verrattuna koko tapausten määrään, jos numerus on suuri. Pienissä otoksissa tällä kuitenkin on merkitystä. Niinpä t- taulukon arvo haetaan vapausasteiden eikä numeruksen kohdalta.

Olemme nyt valmiita toteamaan, oliko miesten ja naisten ero järkeilytestissä merkitsevä vai ei. Periaatteessa voimme tehdä tämän kahdella hieman erilaisella tavalla. Jos me olemme etukäteen päättäneet, minkä kokoisen riskin haluamme ottaa, tutkimme onko saamamme arvo suurempi kuin tämän riskitason raja-arvo. Oletetaan aluksi, että päätimme tehdä testauksen 5 %:n riskitasolla. Tällöin siis toteamme etukäteen, että hyväksymme eron "todeksi", jos sen saaminen sattumalta on mahdollista vain 5 %:ssa tapauksista, tai harvemmin. Haemme taulukon suuntaamatonta testausta (two-tailed test) vastaavan 5 %:n sarakkeen vapausasteiden 15+15-2=28 kohdalla olevan arvon (taulukko on tehty niin, että

tällöin itse asiassa otetaan jakauman molemmista päistä 2.5 %). Tämä on 2.048. Tämä arvo olisi riittänyt, jotta ero olisi ollut merkitsevä 5 %:n riskitasolla. Koska oma arvomme, 2.79, ylittää tämän, teemme johtopäätöksen, että ero on todellinen (eli keskiarvot eivät ole yhtä suuret perusjoukoissa).

Tähän asti olemme verranneet kahden eri ryhmän keskiarvoja, naisten ja miesten, viriketaustaltaan hyvien ja huonojen jne. Kyseessä voi kuitenkin olla myös tilanne, jossa samoilta henkilöiltä on hankittu kaksi eri arvoa. Näin voimme laskea kaksi keskiarvoa ja verrata niitä, vaikka meillä on vain yksi ryhmä. Koska muuttujat saattavat korreloida, pyrkivät olemaan kullakin henkilöllä samansuuntaiset, puhutaan nyt riippuvista keskiarvoista tai korreloivista ryhmien keskiarvoista. Vaikka kyse on keskiarvojen eroista, on itse asiassa kätevintä laskea kullekin henkilölle saatujen mittalukujen ero, d ja sen keskiarvo, ja käyttää seuraavaa kaavaa (periaatekaava oikealla tuottaa tietysti saman tuloksen):

$$t = \frac{\bar{d}}{\sqrt{\frac{\sum d^2 - (\sum d)^2 / N}{N - 1}}} \qquad t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2_{\overline{X}_1} + S^2_{\overline{X}_2} - 2rS_{\overline{X}_1} S_{\overline{X}_2}}}$$

Keskivirhe pienenee, kun alku- ja loppumittaus korreloivat (positiivisesti). Seuraavassa esimerkissä sekä osoittaja että nimittäjä on kerrottu N:llä.

Oletetaan vaikka, että meidän esimerkkiaineistomme 10:lle ensimmäiselle henkilölle on annettu lisäharjoitusta erilaisissa kielellisissä tehtävissä ja testataan hypoteesia, että kielellisen testin tulos paranee harjoituksen jälkeen. Ensimmäinen mittaus on se, joka on aiemmin esitetty, mittaus 2 tehdään harjoituksen jälkeen. Saamme henkilölle seuraavat testitulokset:

koehenkilö	mittaus 1	mittaus 2	d	d^2
1	22	23	—1	1
2	26	28	—2	4
3	29	27	2	4
4	24	24	0	0
5	24	25	—1	1
6	28	29	—1	1
7	23	26	—3	9
8	24	25	—1	1
9	26	25	1	1
10	29	29	0	0
			Σd = —6	$\Sigma d^2 = 22$

Alku- ja loppumittauksen arvojen välille on mahdollista laskea korrelaatio. Se on verraten korkea r= .821. Tietoa voidaan käyttää hyväksi. Sen tietäminen pienentää erotuksen keskivirheen arviointia. Ilman tätä tietoa erotusten keskivirheeksi tulisi 1.034 (kun se nyt on 0.451).

Kunkin henkilön kahden mittatuloksen jälkeen on laskettu arvojen erotus (d) sekä sen neliö (d-toiseen). Sarakkeiden alla on näiden summat. Huomaa, että eroja ja niiden summaa laskettaessa täytyy etumerkit ottaa huomioon, kyseessä eivät ole itseisarvot. Kun sijoitamme luvut kaavaan, saamme seuraavaa:

$$t = \frac{-6}{\sqrt{\frac{10 \cdot 22 - (-6)^2}{10 - 1}}} = \frac{-6}{\sqrt{\frac{220 - 36}{9}}} = \frac{-6}{\sqrt{20.44}} = -1.33$$

Saatu t:n arvo on -1.33. Etumerkki kertoo, että mittausten taso on kasvanut (kun erotukset lasketaan näin). Tulos oli parempi toisella kerralla, oletuksemme mukaisesti. Periaatteessa voisimme tehdä hypoteesin testauksen suunnattuna (yksitahoisena). Sitä käytetään kuitenkin hyvin harvoin. Tässä tapauksessa sillä ei ole merkitystä. Emme voi hylätä nollahypoteesia. Olisimme tarvinneet t:n arvon 2.262 (suuntaamaton, df=9) tai t:n arvon 1.833 (suunnattu testaus), jotta olisimme voineet olettaa tällaisen eron olevan riittävän harvinainen sattumalta tulleena ja olisimme voineet pitää sitä "aitona" erona.
Tunnusluku erotus jaettiin erotuksen keskivirheellä ja todettiin, että kriittinen suhde ei kasva riittävän korkeaksi.

Tässä on hyvä mainita, että t-testi on erityistapaus yksisuuntaista varianssianalyysia (kun ryhmien lukumäärä on 2). Riippuvien mittausten t-testi on taas erityistapaus riippuvien mittausten varianssianalyysista.

b) Kahden prosenttiluvun erotus

Prosenttilukujen eroa testattaessa ovat perusperiaatteet aivan samat kuin keskiarvojen erojen yhteydessä. Meidän on siis laskettava eron keskivirhe ja suhteutettava havaittu ero siihen. Tällä kertaa on jakauma on laadittu vain suuria

otoksia ajatellen, joten on käytettävä normaalijakauman todennäköisyysarvoja, ts. on tehtävä Z-testi. Keskivirheen saamme seuraavasta kaavasta:

$$S_{d_p} = \sqrt{Pe\ Qe\left(\frac{N_1 + N_2}{N_1\ N_2}\right)} \qquad Pe = \frac{N_1\ P_1 + N_2\ P_2}{N_1 + N_2}$$

$$Qe = 100 - Pe$$

Kaava vaikuttaa hiukan monimutkaiselta, koska siihen lasketaan ensin painotettu estimaatti populaation prosenttiluvusta (Pe), mutta laskennallisesti se ei ole sen kummempi kuin aiemmatkaan. Oletetaan, että tutkija on kehittänyt aineen, jonka uskoo lisäävän rottien oppimiskykyä. Sen vaikutusta hän tutkii panemalla rotat sokkeloon, josta niiden tulee selvitä asetetussa määräajassa. Ainetta saaneita rottia on 50 ja niistä selvitti sokkelon 36 %. Vertailuryhmässä on 60 rottaa ja 42% niistä selvisi sokkelosta määräajassa. Ensimmäiseksi on huomattava, että etukäteen paremmaksi kuviteltu ryhmä olikin huonompi, joten emme voi tehdä suunnattua testiä, vaikka olisimme niin ensin ajatelleetkin. Voi jopa esiintyä paradoksaalinen erhe: ero on huomattava ja tavallisin mittapuin tilastollisesti merkitsevä mutta väärään suuntaan (joskus nimitetään kolmostyypin erheeksi). Aloitamme laskemalla Pe- ja Qe-arvot:

$N_1 = 50$
$P_1 = 36$
$N_2 = 60$
$P_2 = 42$

$$Pe = \frac{50 \cdot 36 + 60 \cdot 42}{50 + 60} = 39.27$$

$$Qe = 100 - 39.27 = 60.73$$

Kun nämä arvot sijoitetaan edellä esitettyyn keskivirheen kaavaan, saadaan Sdp=sqrt(2384*0.037)=9.39

Saatu arvo, 9.39 on prosenttilukujen erotuksen keskivirhe, ts. erotuksen otantajakauman standardipoikkeama (hajonta). Prosenttilukujen ero on vielä jaettava tällä, jotta tiedetään, kuinka monen standardipoikkeaman päässä otantajakauman keskeltä ("ei eroa tilanteesta") ollaan. Yhtä hyvin olisi tietysti eron P1 - P2, voinut panna heti osoittajaksi ja edellä lasketun keskivirheen kaavan nimittäjäksi, jolloin tulos olisi ollut suoraan haluttu Z-arvo (kriittinen suhde). Nyt on siis jakolasku vielä jäljellä, saamme:

$$Z = \frac{P_1 - P_2}{S_{d_p}} = \frac{36 - 42}{9.39} = -0.64$$

Saatu ero on siis otantajakaumassa 0.64:n standardipoikkeaman kohdalla. Muistettaneen, että etumerkillä ei ole merkitystä, lukua käytetään itseisarvona. Taulukosta tai aiemmin esitetystä normaalijakauman kriittisiä arvoja kuvaavasta kuvasta voimme todeta, että suuntaamattomassa testauksessa olisimme tarvinneet arvon 1.96, jotta olisimme päässeet edes 5 %:n merkitsevyyteen. Tutkija joutuu toteamaan, että rottien suoritusten erot voivat olla yhtä hyvin sattuma, ero ei ole tilastollisesti merkitsevä.

Kahden %- tai suhdeluvun erotus on tässä perinteisellä tavalla esitettynä. Suhdelukujen erotus jaetaan erotusten keskivirheellä. Tuloksesta päätellään, voisiko oletus nollaerotuksesta jäädä voimaan.

Käytännössä tätä tapaa laskea ei juuri käytetä. %-luvut lasketaan frekvensseistä. Frekvensseihin perustuva ristiintaulukko ja khiin neliö ovat käytännössä se tapa, jolla tarkastelu tehdään.

c) Kahden korrelaatiokertoimen erotus

Korrelaatiokertoimet voivat olla periaatteessa kahdenlaisessa suhteessa toisiinsa: riippuvia tai riippumattomia. Kun kertoimet ovat riippuvia, on tavallisesti laskettu samalla koehenkilöjoukolla yhden muuttujan korrelaatiot muihin. Niinpä meidän esimerkkiaineistossamme ovat esim. viriketaustan korrelaatiot muihin muuttujiin riippuvia. Tämäkin voidaan tutkia: ovatko kaksi korrelaatiota, joissa toinen muuttuja on sama saman suuruisia vai poikkeavatko ne toisistaan (riippuvat korrelaatiot).

Tavallisin ja selvin tapaus riippumattomista korrelaatioista on se, jossa kahdelle eri henkilö- joukolle on laskettu samojen muuttujien korrelaatiot. Esimerkiksi meidän aineistossamme on viriketaustan ja järkeilytestin korrelaatio miesten joukossa .63 ja naisten joukossa .40. Tällaiset riippumattomat korrelaatiot ovat tavallisin ja tärkein alue, jossa erotuksen merkitsevyyttä tarvitaan ja sitä käsitellään tässä lähemmin. Kaava on seuraava:

$$Z = \frac{Z_{F_1} - Z_{F_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

Kaavassa esiintyy kahdenlaisia Z-symboleja. Ensimmäinen, joka siis ilmaisee, että laskutoimituksesta saadaan tulokseksi Z-arvo, tarkoittaa sitä, että otantajakauma oletetaan on normaaliksi ja tulos on normaalijakauman standardipisteitä, Z-pisteitä. Osoittajassa olevat Z_F-symbolit taas edustavat Fisherin Z-pisteitä, joihin tutustuimme korrelaation luottamusväliä laskettaessa. Niiden tarkoituksenahan on saattaa korrelaatioiden todennäköisyyksiä kuvaavat vinot jakaumat normaaleiksi. Tutkittavat korrelaatiot, .63 ja .40 on siis ensin muutettava Fisherin Z-pisteiksi, jonka jälkeen voidaan suorittaa lasku kaavan mukaan. Muunnetut arvot ovat .741 ja .424. Miehiä ja naisia on aineistossa molempia 15, joten saamme laskutoimituksen tuloksena erotukselle 0.317 keskivirheen 0.408 ja siitä Z-suhteen 0.776. Olemme siis vain jonkin matkaa sattumaodotuksesta, jolloin se on vielä varsin mahdollinen selitysperuste.

Toisin kuin luottamusväliä laskettaessa, jolloin tulos oli ala- ja yläraja korrelaatioina, ei nyt ole mitään tarvetta muuttaa saatua Z-arvoa takaisin korrelaatioiksi. Saatu tulos on eron sijainti normaalijakaumassa ja voidaan käyttää sellaisenaan. Voidaan helposti todeta, että saatu arvo on niin pieni, ettei se yllä millekään normaalisti käytössä olevalle merkitsevyystasolle. Teemme siis johtopäätöksen, että viriketaustan ja järkeilytestin korrelaatio on sattuman aiheuttaman virheen puitteissa sama miehillä ja naisilla, ero ei ole merkitsevä.

6. Korrelaatiokertoimen merkitsevyys

Paitsi kahden korrelaation eron suuruutta, voidaan tarkastella myös yhden korrelaation merkitsevyyttä sellaisenaan. Tällöin kysytään, poikkeaako se riittävästi nollasta ts. onko korrelaatio riittävän suuri ollakseen "todellinen". Jonkin kokoisia korrelaatioitahan voi saada pelkästään sattumalta sellaisestakin populaatiosta, jossa korrelaatio on nolla. Korrelaation merkitsevyyttä voidaan tutkia seuraavalla kaavalla (korrelaatio jaettuna keskivirheellä, df=N-2):

$$t = \frac{r}{\sqrt{1 - r^2} / \sqrt{N - 2}}$$

Tässä tapauksessa ei ole tarvetta Fisherin Z-muunnoksiin, koska oletetun nollan nollan kohdalla korrelaation otantajakauma on symmetrinen. Voimme vaikkapa tutkia, onko sukupuolen ja järkeilytestin tuloksen korrelaatio, .464, merkitsevä. Saamme t:n arvoksi 2.77.

Yhden korrelaation merkitsevyyttä testattaessa on vapausasteiden määrä N-2. Katsomme siis t-taulukkoa riviltä 28, sarakkeen valitsemme suuntaamattoman testin tapaan. Saamamme arvo t=2.77 riittää niukasti 1 %:n merkitsevyystasolle. Voimme siis todeta, että sukupuolen ja järkeilytestin välinen yhteys on tilastollisesti merkitsevä.

Tässä on haluttu käyttää muuttujaa sukupuoli, joka on dikotominen. Onko tulomomenttikerrointa oikeutettua käyttää tässä tapauksessa. Vastaus on kyllä. Kvalitatiivinen kaksiluokkainen muuttuja voi olla toisena muuttujana.

Korrelaation merkitsevyys voidaan todeta suoraan taulukosta ilman laskutoimituksia (taulukko 4). Tulos tulee suoraan tietokoneohjelmissa ja t-jakauman tarkka p-arvo on ilmoitettu. Kun se menee pieneksi (p=.05 tai pienempi) hylkäämme korrelaatiota koskevan nollahypoteesin.

7. Useamman kuin yhden eron yhtäaikainen testaus

Silloin tällöin tutkijalla on kokonainen joukko tunnuslukuja, joiden välisiä merkitsevyyttä pitäisi tutkia. Osa analyyseista on syntynyt kokeellisessa tutkimusympäristössä. Tällainen tilanne on esimerkiksi kasvien viljelykeinojen kehittäjällä. Hänellä saattaa olla vaikkapa viidessä eri lämpötilassa kasvatettuja taimia. Samaan aikaan saattavat lannoitteet vaihdella siten, että jossakin koeruudussa on vaikkapa kylmässä kasvaneita, vähän lannoitettuja taimia, toisessa voi olla kylmässä kasvaneita, paljon lannoitettuja ja kolmannessa lämpimässä kasvatettuja, vähän lannoitettuja taimia jne. Onko nyt lämpötilalla vaikutusta kasvuun? Miten lannoitus vaikuttaa ja miten lämpötila ja lannoitus yhdessä? Näiden asioiden jäljille pääsee kätevimmin testaamalla yhdellä kertaa, onko koko joukossa merkitseviä eroja vai ei.

Yleensä kyseessä ovat joko frekvenssit tai keskiarvot. Jos vaikka annamme jonkin joukon jakaantua ryhmiin sen mukaan, mitä ehdotusta he kokouksessa kannattavat, saamme ryhmiä, joiden frekvenssit voivat olla lähellä tai kaukana toisistaan. Toisin sanoen, voi olla, että jakaantuminen on niin lähellä tasajakoa, ettei ryhmien voi katsoa olevan eri suuruisia, tai sitten jakaantuminen on niin epätasaista, että erot tuskin voivat olla pelkkää sattumaa. Tällaiseen usean ryhmän frekvenssien erojen merkitsevyyteen sopii khiin neliö-testi.

Toisessa tapauksessa voisimme vaikka verrata keskimääräistä elokuvissakäynnin, määrää eri kaupunginosien kesken. Nyt meillä olisi joukko keskiarvoja, jotka voivat olla lähellä tai kaukana toisistaan. Näiden erojen yhtaikaisessa merkitsevyystestauksessa käytetään varianssianalyysia.

a) khiin neliö

Alkuosassa on jo alustavasti käsitelty ristiintaulukkoa ja khiin neliön laskemista varsin pitkälle.

Khiin neliötä voimme käyttää havaittujen arvojen ja teoreettisten arvojen erotuksen tutkimiseen. Esitämme ensin tilanteen, jossa teoreettiset arvot eivät ole

riippumattomuuslukuja vaan teoreettisen jakauman odotuksia. Toiseksi tulee tavanomainen 2 x 2 -ristiintaulukko. Palaa sitten tarkastelemaan kuvaavassa osassa esitettyä ristiintaulukkoa ja khiin neliötä 3 x 3 -ristiintaulukosta.

Eläinten haluamia elinolosuhteita tutkittaessa voidaan järjestää tila, jossa eläimet saavat vapaasti hakeutua siihen osaan, joka parhaiten vastaa niiden tarpeita. Ajatellaan vaikkapa, että sata hiirtä saa vapaasti valita elinpaikkansa laatikostosta, jossa on neljä eri lämpöistä osaa. Riittävän ajan kuluttua tutkitaan, miten monta hiirtä kussakin osassa on. Jakauma osoittautuu olevan 10, 28, 32 ja 30. Khiin neliö-testillä voidaan tutkia, poikkeaako jakauma merkitsevästi siitä, mitä saataisiin täysin sattumanvaraisesti.

Khiin neliön oleellinen ajatus on juuri kahden jakauman vertailu: sen, joka saataisiin, jos jakauma olisi sattumanvarainen, jos siihen ei vaikut taisi mikään systemaattinen tekijä sekä sen, joka kokeessa tai muussa tutkimuksessa on saatu. Saamme siis kahdenlaisia frekvenssejä, havaittuja (fo, observed frequenqy) sekä odotettuja (fe, expected frequency). Tässä esimerkissä ovat odotetut frekvenssit yksinkertaisesti 100:4 = 25, koska tasajako neljään ryhmään sadasta tuottaa kahdenkymmenenviiden ryhmät. Meillä on siis seuraavat havaitut ja odotetut frekvenssit:

fo	10	28	32	30
fe	25	25	25	25

Khiin neliön laskukaava (joka on jo esiintynyt kontingenssikertoimen ja ristiintaulukoinnin yhteydessä), on seuraava:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Havaittujen ja odotettujen frekvenssien erot siis korotetaan toiseen, jaetaan odotetuilla frekvensseillä ja lopuksi kaikki lasketaan yhteen. Saamme seuraavat arvot: 225/25 + 9/25 + 49/25 + 25/25 = 13.32

Khin neliön arvo on siis 12.32, mutta taulukon käyttöä varten tarvitsemme vielä vapausasteet. Yksisuuntaisessa tapauksessa, jossa meillä siis on vain yksi vaikuttava tekijä (tässä: lämpötila) ja saamme havaituista frekvensseistä yksisuuntaisen jakauman (emmekä ristiintaulukkoa), on vapausasteiden määrä yhtä vähemmän kuin jakaumassa on luokkia, siis 4-1=3. Saatu arvo ylittää taulukossa kolmen vapausasteen rivillä olevan arvon 11.3, joka on yhden prosentin merkitsevyyden raja. Toteamme siis, että lämpötilat jakoivat hiiret erikokoisiin ryhmiin; riski, että näin tapahtuikin sattumalta, on tällöin pienempi kuin 1 %.

Kaksiulotteisessa tapauksessa (ristiintaulukko) on periaate aivan sama, mutta odotetut frekvenssit hankitaan reunajakaumien avulla. Tämän teimme jo kontingenssikertoimen yhteydessä laskiessamme asuinpaikan laadun ja kolmen eri kahvimerkin suosion välistä yhteyttä. Khiin neliön arvoksi saimme 20.03. Kaksiulotteisessa tapauksessa on vapausasteiden määrä (r-l)*(k-l) eli rivien määrä vähennettynä yhdellä kertaa sarakkeiden määrä vähennettynä yhdellä, tässä siis (3-1)*(3 -1)=4.

Taulukosta toteamme, että saatu arvo ylittää vielä 0.1 %:n rajankin, joka on 18.5. Asuinpaikan laadun ja kahvilaadun välillä on siis erittäin merkitsevä yhteys.

Khiin neliötä laskettaessa ovat lähtöluvut aina frekvenssejä. Jos tulos annetaan vaikkapa prosentteina tai suhdelukuina, on ne ensin muutettava frekvensseiksi, muuten kaava ei toimi oikein. Frekvenssit ovat tyypiltään tapaus epäjatkuvasta muuttujasta, ne ovat aina kokonaislukuja, ne eivät koskaan saa arvoja niiden väliltä. Khiin neliön taulukkoarvot on kuitenkin laskettu ikään kuin muuttuja olisi jatkuva. Teoreettiset arvot ovat taas aina tarkkoja ja sisältävät desimaaliosan.

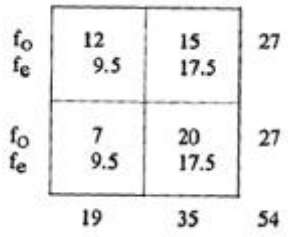
Havaittujen arvojen epäjatkuvuus voidaan erityisesti 2*2 -taulukossa korjata. Tällöin käytetään ns. Yatesin korjausta, jossa havaittujen ja odotettujen frekvenssien eron itseisarvosta vähennetään arvo 0.5. Kaava saa tällöin seuraavan muodon:

$$\chi^2 = \Sigma\left(\frac{\left(\left|f_o - f_e\right| - 0.5\right)}{f_e}\right)$$

Täsmällistä rajaa siihen, milloin korjausta tulisi käyttää, ei ole, mutta seuraavat linjat ovat suositeltavia:

- Kun vapauasteita on 1.
- Kun enemmän kuin 20 % fe-arvosta on alle 5
- Kun numerus on alle 40 ja yksikin fe-arvo on alle yhden.

Kuvitellaan saaduksi seuraava 2 x 2 -taulukko (nelikenttä):



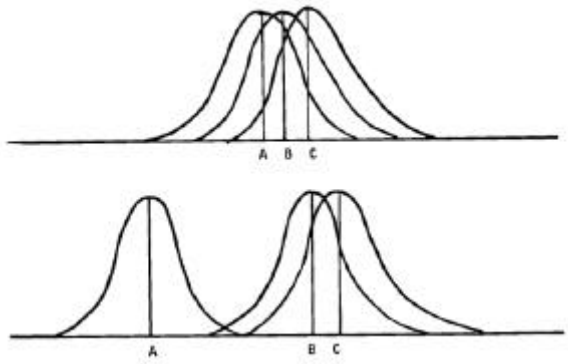
Vapausasteita on nyt (2-1)*(2-1)=1, joten käytämme Yatesin korjausta. Saamme seuraavan laskutoimitusten perusteella soluihin arvot:

Khii-toiseen = 0.42 + 0.23 + 0.42 + 0.23 = 1.3

Saatu arvo jää alle 5 %:n tason, johon olisi vaadittu khiin neliö 3.8, joten toteamme, että muuttujien välillä ei ole merkitsevää yhteyttä. Yatesin korjauksena esitetylle tavalle on edellä annettu paljon tilaa. Se viittaa yleisempään ongelmaan. Ristiintaulukoinnin yhteydessä solukko pitäisi täyttyä siten, että odotusarvot eivät muodostu kovin pieniksi (alle 5). Tiheätä luokittelua täytyy usein karkeistaa. On hyvät tapa katsoa ristiintaulukkoa sekä frekvensseinä (fo) että odotusarvoina (fe). Tällöin havaitsee sen, milloin joku tai jotkut fe-arvot menevät kovin pieneksi. Samoin löytää ne solut, joissa nämä arvot eniten poikkeavat toisistaan. Vertailu helpottaa tulosten esille saamista.
Khiin neliö kertoo onko eroja. Vasta vertailu havaittujen ja odotettujen arvojen välillä kertoo missä erot sijaitsevat.

b) Varianssianalyysi

Totesimme varianssianalyysin olevan keino tutkia saman aikaisesti useiden keskiarvojen erojen merkitsevyyttä. Perusperiaate on aivan sama kuin se, joka esitettiin jo eta-kertoimen yhteydessä: vaihtelun jakaminen ryhmien sisäiseen ja ryhmien väliseen komponenttiin. Tarkastellaanpa seuraavia tapauksia, joissa on molemmissa kolmen ryhmän jakaumat piirrettyinä samalle asteikolle:



Ylemmässä kuvassa ovat jakaumien keskiarvot varsin lähellä toisiaan; ryhmien välillä ei ole paljoakaan eroa. Saman asian voimme sanoa hieman teknisemmin toteamalla, että tässä tapauksessa muodostuu kokonaisvaihtelu pääasiassa ryhmien sisäisestä vaihtelusta, ryhmien välinen vaihtelu on vähäistä.
Alemmassa kuvassa on ryhmien välillä huomattavasti suuremmat erot, tarkemmin sanoen A:n erot B:hen ja C:hen ovat melko suuret, B:n ja C:n ero ei ole suurempi kuin ylemmässäkään tapauksessa. Joka tapauksessa voidaan sanoa, että verrattuna ylempään tapaukseen kokonaisvaihtelusta on paljon suurempi osa nyt ryhmien välistä vaihtelua. Juuri tästä on varianssianalyysissä kysymys: jos riittävän suuri osa kokonaisvaihtelusta on ryhmien välistä, katsotaan, että ryhmien välillä on merkitseviä eroja. Tämä voidaan sanoa niinkin, että tällaisessa tapauksessa on epätodennäköistä, että ryhmät edustaisivat samaa populaatiota.

Eta-kertoimen yhteydestä muistettaneen, että kokonaisneliösumma (SStotal) voidaan jakaa ryhmien sisäiseen neliösummaan (SSwithin) ja ryhmien väliseen neliösummaan (SSbetween). Mitä suurempi on ryhmien välisen vaihtelun osuus, sitä merkitsevämpiä ovat erot. Varianssianalyysissä ei neliösummia verrata

suoraan toisiinsa, vaan ne muutetaan ensin varianssiestimaateiksi jakamalla ne vastaavilla vapausasteilla. Ryhmien välisen varianssiestimaatin suhde ryhmien sisäiseen, ns. F-suhde, on varianssianalyysin keskeinen tulos, jonka merkitsevyys nähdään taulukosta.

Laskemme tässä neliösummat hieman eri tavoin kuin eta-kertoimen yhteydessä, tulos on tietenkin sama. Tarvitsemme seuraavat kaavat.

Kokonaisneliösummavaihtelu:

$$SS_t = \sum X - \frac{\left(\sum X\right)^2}{N}$$

Tämähän on sama kuin (N-1)*varianssi, kun mitään ryhmiin kuulumista ei oteta huomioon. Vapausasteet ovat N-1.

Kun ryhmäerot poistetaan ja lasketaan sisäisen vaihtelun neliösumma:

$$SS_w = \sum \left(\sum X_g^2 - \frac{\left(\sum X_g\right)^2}{N_g} \right)$$

Tehdään SS-laskelma kussakin ryhmässä ja lasketaan näin saadut komponentit yhteen. Keskiarvojen vaihtelu ryhmien välillä on mitätöity. Vapausasteet ovat N-g, kun g on ryhmien lukumäärä. Ng on kunkin osaryhmän numerus.

Kun vaihteluun otetaan mukaan vain keskiarvojen erot mutta ei sisäistä vaihtelua lasketaan:

$$SS_b = \sum \left(\frac{\left(\sum X_g\right)^2}{N_g} \right) - \frac{\left(\sum X\right)^2}{N}$$

Siinä on samoja osatekijöitä kuin aikaisemmissa kaavoissa. Vapausasteet ovat g-1.

Itse asiassa kaikkia kolmea ei ole tarpeen laskea, koska kokonaisneliösumma on ryhmien välisen ja sisäisen neliösumman summa. Kun kaksi näistä tiedetään,

voidaan kolmas laskea. Kaikki kolme voidaan kuitenkin laskea, jolloin saadaan myös tarkistusmahdollisuus: välisen ja sisäisen summan täytyy olla kokonaisneliösumma. Voimme esimerkkinä tutkia, onko aineistossamme merkitsevää eroa viriketaustan mukaan jaettujen ryhmien matematiikan numeroiden keskiarvoissa. Keskiarvot ovat 6.67, 7.15 ja 8.00. Kaavoissa esiintyvät termit SummaX (pisteiden summa) sekä SummaX^2 (toiseen korotettujen pisteiden summa). Nämä saamme parhaiten näkyville taulukoimalla pisteet viriketaustan mukaisiin ryhmiin ja laskemalla sarake jokaisen pistearvon neliöille. Sarakesummat ovat tällöin edellä mainitut termit. Saamme seuraavan taulukon:

Viriketausta					
huono		keskinkertainen		hyvä	
X	X^2	X	X^2	X	X^2
8	64	7	49	5	25
7	49	9	81	8	64
6	36	7	49	9	81
6	36	7	49	8	64
7	49	7	49	10	100
6	36	8	64	9	81
6	36	7	49	8	64
7	49	6	36	7	49
7	49	8	64		
		7	49		
		7	49		
		6	36		
		7	49		
Summa 60	404	93	673	64	528

Kokonaisneliösummaa laskettaessa tarvitaan kaikkien pistearvojen summaa (SummaX), joka korotetaan toiseen ja jaetaan tapausten kokonais- määrällä (N). Tämä vähennetään kaikkien toiseen korotettujen pistearvojen summasta (SummaX^2). Laskutoimitus on seuraava:

$$SS_t = (404 + 673 + 528) - \frac{(60 + 93 + 64)^2}{9 + 13 + 8}$$

$$= 1605 - \frac{47089}{30} = 35.37$$

Ryhmien sisäistä neliösummaa laskettaessa tarvitaan ryhmittäiset pistearvojen summat (SummaXg), jotka korotetaan toiseen ja jaetaan vastaavien ryhmien tapausten määrillä (Ng). Tulokset vähennetään ryhmittäisten toiseen korotettujen pistearvojen summista (SummaX^2g). Kunkin ryhmän saamat arvot lasketaan yhteen. Saamme osaryhmistä arvot 4.00, 7.69 ja 16.00, joiden summa 27.69 on SSw.

Ryhmien välisen neliösumman saa vähentämällä: 35.37-27.69=7.68. Voimme kuitenkin tarkistusmielessä laskea senkin erikseen. Lasketaan kullekin ryhmälle pistearvojen summa, korotetaan se toiseen ja jaetaan ryhmän numeruksella. Nämä lasketaan yhteen ja summasta vähennetään kokonaisnumeruksella jaettu, toiseen korotettu kaikkien pistearvojen summa. Tulos on noin 7.67.

Ryhmien välinen neliösumma ja ryhmien sisäinen neliösumma oli vielä muutettava keskineliöksi (MS, mean square) jakamalla ne vastaavilla vapausasteilla. Näissä luvuissa on hyvin keskeinen asia. Kun SS-tesmit jaetaan vapausasteillaan, niistä saadaan perusjoukon hajonnan arvioita. Jos välisestä vaihtelusta saatu arvio on huomattavasti suurempi kuin sisäisestä vaihtelusta saatu, ryhdymme epäröimään: voiko näin suuri keskiarvojen vaihtelu olla pelkkää otantasattumaa. Tulosten suhde on F-arvo:

$$MS_b = \frac{SS_b}{g-1} = \frac{7.67}{3-1} = 3.84$$

$$MS_w = \frac{SS_w}{N-g} = \frac{27.69}{30-3} = 1.03$$

$$F = \frac{MS_b}{MS_w} = \frac{3.84}{1.03} = 3.73$$

Varianssianalyysissa ei siis neliösummavaihtelua verrata suoraan toisiinsa. Neliösummista meillä on mahdollisuus laatia kaksi estimaattia perusjoukon varianssille. Ensimmäinen (MSb) perustuu keskiarvojen vaihteluun otosten/ryhmien välillä. Toinen (MSw) arvio taas pohjautuu ryhmien sisäiseen vaihteluun. Kahden varianssiestimaatin suhde, F-suhde, muodostaa otantajakaumana F-jakauman. Sen muoto määrittyy kahdesta vapausasteesta: osoittajan ja nimittäjän. F-jakauman mukaan arvioidaan se kasvaako välinen vaihtelu liian isoksi ollakseen pelkästään sattumaa.

F-suhteen merkitsevyyttä taulukosta katsottaessa on otettava huomioon osoittajan ja nimittäjän vapausasteet erikseen. Tässä tapauksessahan on osoittajaa vastaavia vapausasteita g-l=2 ja nimittäjän kohdalla N-g=27. Taulukosta haemme osoittajan vapausasteiden mukaisen sarakkeen, siis toisen, ja nimittäjän vapausasteiden mukaisen rivin numero 27. Huomaamme, että saamamme arvo yltää 5 %:a vastaavaan arvoon 3.35. Viriketaustan mukaan jaettujen ryhmien

matematiikan numeroiden keskiarvoissa on siis eroja 5 %:n riskitasolla. Tämän jälkeen suoritetaan erojen lähempi tarkastelu.

Varianssianalyysin voi tehdä myös useampisuuntaisena. Kaksisuuntainen analyysi on varsin tavallinen, joten käymme sen tässä läpi menemättä kuitenkaan itse laskutoimituksiin, jotka yleensä tehdään valmiilla ohjelmilla. Kaksisuuntaisesta analyysistä on kysymys silloin, kun aineisto on jaettu yhtaikaa kahden muuttujan mukaisiin ryhmiin ja kullekin ryhmälle on laskettu keskiarvot, joiden eroja siis testataan. Tällainen olisi tilanne esim. silloin, jos laskisimme matematiikan numeroiden keskiarvot sukupuolten ja viriketaustan mukaan jaetulle aineistolle.

Useampisuuntainen varianssianalyysi rinnastuu suoraan regressioanalyysiin. Voimme selittää matematiikan keskiarvoa yhtä aikaa sukupuolella ja viriketaustalla. Varianssianalyysissa viriketausta-muuttujaa ei käsitellä kvantitatiivisena vaan kvalitatiivisena. (Regressioanalyysissa se pitäisi purkaa kahdeksi dummy-koodatuksi muuttujaksi.) Varianssianalyysissä saadaan lisäksi muuttujien yhdysvaikutus (interaktio), mitä regressioanalyysissa ei normaalisti saada. Erillisten ja yhteisten selitysosuuksien lisäksi muuttujilla voi olla yhdistelmään perustuvia yhdysvaikutuksia. Näitä voi parin muuttujan tapauksessa tarkastella korrelatiivisessakin tutkimuksessa. Kun selittäviä muuttujia on kolme tai enemmän niiden määrä kasvaa hallitsemattomaksi.

Voimme nyt siirtyä tarkastelemaan, minkälainen tulos kaksisuuntaisesta varianssianalyysistä meidän esimerkkiaineistossamme tulee, kun suuntina ovat viriketausta ja sukupuoli sekä testattavina matematiikan numerot. Saamme seuraavan taulukon, johon on merkitty ryhmittäiset keskiarvot, hajonnat ja numerukset:

Descriptive Statistics

Dependent Variable: V6

V1	V2	Mean	Std. Deviation	N
nainen	koyha	6.75	.50	4
	keskinkertainen	6.71	.49	7
	rikas	7.25	1.71	4
	Total	6.87	.92	15
mies	koyha	6.60	.89	5
	keskinkertainen	7.67	.82	6
	rikas	8.75	.96	4
	Total	7.60	1.18	15
Total	koyha	6.67	.71	9
	keskinkertainen	7.15	.80	13
	rikas	8.00	1.51	8
	Total	7.23	1.10	30

Taulukkoa silmämäärin tarkastellessa huomaa, että siinä on ainakin jonkin verran sekä rivien välistä, sarakkeiden välistä että interaktiovaikutusta: alemman rivin arvot ovat suurempia kuin ylemmän, sarakkeiden arvot kasvavat vasemmalta oikealle. Interaktio näkyy siinä, että viriketaustan vaikutus näkyy miesten ryhmässä selvemmin. Erilaiset varianssianalyysiohjelmat tulostavat hieman eri tavoin, mutta oleellisin käy ilmi seuraavasta:

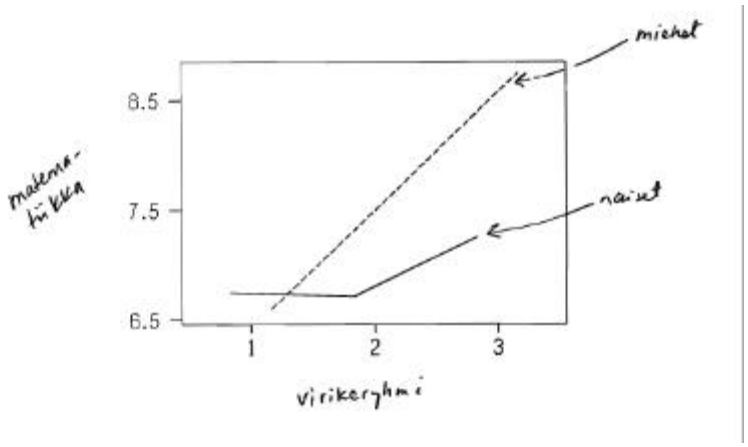
Tests of Between-Subjects Effects

Dependent Variable: V6

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15.155[a]	5	3.031	3.599	.014
Intercept	1518.3	1	1518.349	1802.9	.000
V1	4.209	1	4.209	4.997	.035
V2	7.484	2	3.742	4.443	.023
V1 * V2	3.054	2	1.527	1.813	.185
Error	20.212	24	.842		
Total	1605.0	30			
Corrected Total	35.367	29			

a. R Squared = .429 (Adjusted R Squared = .309)

Tuloksen mukaan voidaan todeta, että keskiarvoissa on sp-ryhmien välillä sattuman rajat ylittävää vaihtelua. Naiset ovat menestyneet heikommin kuin miehet (kun virike ja yhdysvaikutus vakioitu). Virikeryhmien välillä on myös systemaattista vaihtelua. Mitä rikkaampi viriketausta sitä korkeampi matematiikan arvosana. Yhdysvaikutusta ei ole. Silmään kyllä sattuu se, että miehillä viriketaustan yhteys on voimakkaampi, mutta merkitseväksi tulos ei yllä. Ilmiö on graafisesti ohessa:



Miehillä oleva yhteys on erilainen kuin naisilla. Pieni otoskoko vaikeuttaa tilanteen arvioimista. Yhdysvaikutus näkyy, mutta se ei kohoa tilastollisesti merkitsevälle tasolle.

Eta-toiseen kertoimet ilmaisevat yhteyden selvyyttä. Sukupuolen osalta se on 11.4 % (neliösummista laskettu SSsp/SStot). Viriketausta selittää 22.8 %. Yhdysvaikutuksen osuus on 8.6 %. Kokonaisselitys on 42.9 % (yksi miinus SSerror/SStot). Kokonaisselitys ei ole osaselitysten summa (vaikka siltä näyttäisi tässä), koska sukupuoli ja viriketausta "korreloivat" koska asetelma ei ole ortogonaalinen solufrekvenssien ollessa erisuuret. Huomaa, että Spss-ohjelman ilmoittamat partial eta-toiseen -kertoimet lasketaan eri tavalla. Niitä ei ole tässä tulostuksessa mukana.

Ohessa näkyy eräs tilastollisen tulostuksen piirre. Ohjelmat tuottavat varsin runsaan tulostuksen. Käyttäjän pitää pystyä lukemaan sitä ja tulkitsemaan sen sisältö. Tutkimusraporttiin oheinen tulostus ei sellaisenaan kelpaa. Sitä täytyy muokata lukijalle sopivampaan muotoon. Lukemalla julkaistuja tutkimustekstejä voi arvioida, mikä on sopiva raportointitapa.

Huomautus: kirjaan on jäänyt paikoin epäjohdonmukaisuutta. Kun SS-termi jaetaan N-1:llä saadaan varianssi ja sen neliöjuurena hajonta. Asia pätee sekä muuttujalle yli ryhmien tai ryhmäkohtaisiin arvoihin. Vastaavasti taas hajontatoiseen kerrottuna N-1:llä muuttaa arvot neliösummiksi. Tässä esityksessä on paikoin kuitenkin lukuarvojen pohjana hajonnat ja varianssit, jotka on saatu jakamalla N:llä. Esitys pyritään yhdenmukaistamaan näiltä osin.

Nykyisten tietokoneohjelmien hajonnan tulostus noudattaa N-1 -periaatetta.

8. Testauksen virhetyypit ja efektin suuruus

Empiiristä tutkimusprosessia kuvataan usein kehänä tai spiraalina, joka alkaa teoriaan ja käytäntöön liittyvästä ongelmanasettelusta. Useiden tutkimusvaiheiden kautta päädytään tulosten tulkintaan ja sitä tietä jälleen teoreettisen pohdinnan abstraktille tasolle. Tilastollisten hypoteesien testaaminen on vain pieni osa tässä kehässä. Kontrollin asteeltaan löyhemmässä korrelatiivisessa tutkimusotteessa menetellään tavallisesti siten, että vain todetaan poikkeaako tulos sattumanvaraisesta ja kuinka selvä yhteys on. Kokeellisessa tutkimustraditiossa tilastollinen merkitsevyyden testaus on keskeisemmässä asemassa. Tutkimusasetelman suunnittelu tähtää tilastollisen tarkastelun mahdollisimman häiriöttömään suorittamiseen ja kausaalipäätelmien tekemiseen jopa tulosten yleistettävyyden kustannuksella.

Tilastollisen merkitsevyystestauksen vaiheet voidaan hahmottaa myös seuraavasti:

Perinteisesti tilastollinen päätöksenteko koskee nollahypoteesin (sattumaoletus) hyväksymistä tai hylkäämistä. Kun nollahypoteesi osoittautuu riittävän epäuskottavaksi hylätään se. Tuloksen sanotaan tällöin olevan tilastollisesti merkitsevä tietyllä riskitasolla. On syytä olla selvillä, että päätöksentekoon sisältyy toinenkin riski. Seuraava esitys pyrkii kuvaamaan näitä kahta virhetyyppiä:

Tilanne tuntemattomassa "todellisuudessa"	Empiiristen tulosten perusteella tehty päätös H_0:sta: Hylätään	Hyväksytään
H_0 tosi	Tyypin I virhe Hylkäämisvirhe Alfa-tyypin virhe riski $p = \alpha$	Oikea päätös todennäköisyys $p = 1 - \alpha$
H_0 epätosi	Oikea päätös (ns. voimakkuus) todennäköisyys $p = 1 - \beta$	Tyypin II virhe Hyväksymisvirhe Beta-tyypin virhe riski $p = \beta$

Beta-tyypin virheeseen saa tuntuman, kun ajattelee tilannetta, jossa todellisuudessa muuttujien välillä vallitsee huomattavakin yhteys (esim. korrelaatio), mutta hyvin pieniä otoksia käytettäessä ilmiötä ei saada tilastollisesti merkitsevästi esille. Hyväksymme tällöin sattumamallin (eli nollahypoteesin) vaikka se ei pidä paikkaansa. Tilastollisen testauksen herkkyyttä havaita olemassa olevia seikkoja nimitetään testin voimakkuudeksi.

Viime aikainen suuntaus empiirisissä tutkimuksissa on pyrkimys irrallisista tilastollisista hypoteeseista kohti hypoteesien joukkoa, jotka muodostavat toisiinsa liittyvän, mielekkään kokonaisuuden. Tällaista hypoteesiyhdistelmää voidaan nimittää malliksi. Testauksen kohteeksi nousee mallin ja kokemusperäisesti hankitun aineiston välinen yhteensopivuus. Tällaisia tarkastelutapoja ovat esim. polkuanalyysi ja log-lineaariset menetelmät. Mallia testattaessa ei pyritä niinkään mallin hylkäämiseen vaan sen kehittämiseen ja sovittamiseen aineistoa mahdollisimman hyvin kuvaavaksi. Tutkija pyrkii hyväksymään asetetun hypoteesin eli mallin. Tällaisessa tutkimustilanteessa tavanomaisen riskitason käsitteen rinnalle nousee beta-tyypin virhe ja sen riski. Riittävän suurilla numeruksilla pyritään siihen, että voidaan osoittaa mallin sopivuus aineistoon eli hyväksyä hypoteesi.

Koska tuloksen selvyys (yhteyden voimakkuus) ja tilastollinen merkitsevyys riippuvat otoskoosta, suositellaan nykyisin molempien ilmaisemista. Lukija voi arvioida itse tilannetta. Yhteyden selkeyttä kuvaavia indeksejä ovat esim. korrelaatiokerroin tai sen neliö, yhteiskorrelaation neliö, eta-toiseen.

ES-indeksi (effect size) on alun perin kokeellisen tutkimuksen piiristä. Sen avulla on kuitenkin hyvä havainnollistaa tilastollisen päättelyn kokonaisuutta virhetyyppeineen.

5-portainen skaala on tuttu asennemittauksesta. Sillä tapaa mitattujen osioiden hajontakin on usein lähellä ykköstä. Ajatellaanpa, että koulussa A viihtyminen on saanut keskiarvo 2.75 ja hajonnan 1 ja tutkittujen N=50 ja koulussa B vastaavasti 3.25, hajonta 1 ja N=50. Koulujen ero on ES-mitalla ilmaisten 0.5 (erotus jaettuna verrokkiryhmän hajonnalla alun perin). Pienillä ES-arvoilla ES/2 on sama kuin korrelaatio ryhmään kuulumisen ja viihtymisen välillä.
Voisimme todeta, että pitäisimme yhteyttä merkitsevänä. Se lähestyy jopa 1 %:n rajaa (ollen p=.014), mutta ei aivan yllä siihen.

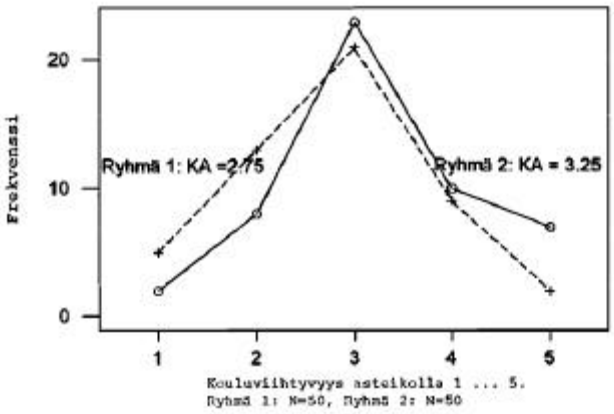
Seuraavaksi vain kuvatut vaiheet:

1) otosten jakaumat, joiden välillä ES=0.5 eli rpbis noin .25
2) erotusten otantajaukauma (kun oletetaan Ho=tosi) ja havaitun arvon sijoittuminen siihen. Se näyttäisi olevan peräisin oletetusta otannasta.
3) voimakkuuden arviointi. Ho=nolla vastaan Hs=on olemassa .5 ero (Hs=spesifi vaihtoehtoinen hypoteesi). Kuinka usein (jos Hs on tosi) tulisimme hyväksyneeksi Ho:n? Tämä on beta-tyypin erhe. Voimakkuus on sen vastatapahtuma eli voimakkuus=1-beta.

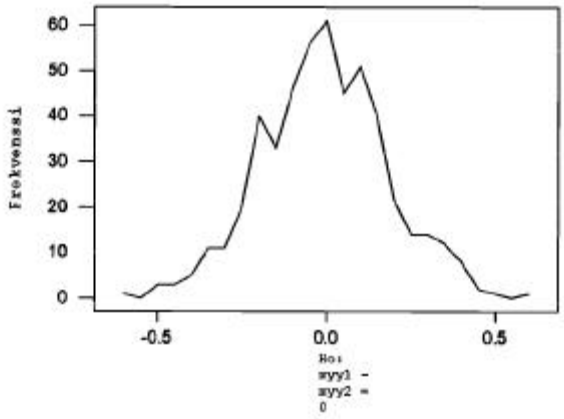
Voimme todeta:

Kun N1 ja N2 ovat molemmat 50 ja hajonnat 1 ja alfa-tyypin erhe on kaksitahoinen .05, niin menettelyn voimakkuus on n. .70. Eli mikäli Hs olisi tosi, löytyisi se tällä menettelyllä tuolla todennäköisyydellä (pidemmän päälle).
Asia on verraten monimutkainen, mutta ei suinkaan mahdoton ymmärtää.

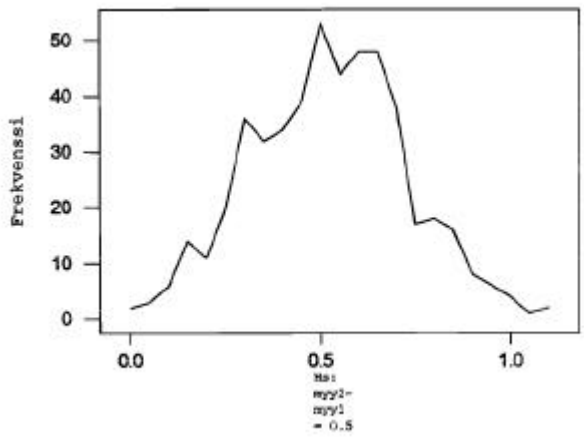
1) Otosten jakaumat:



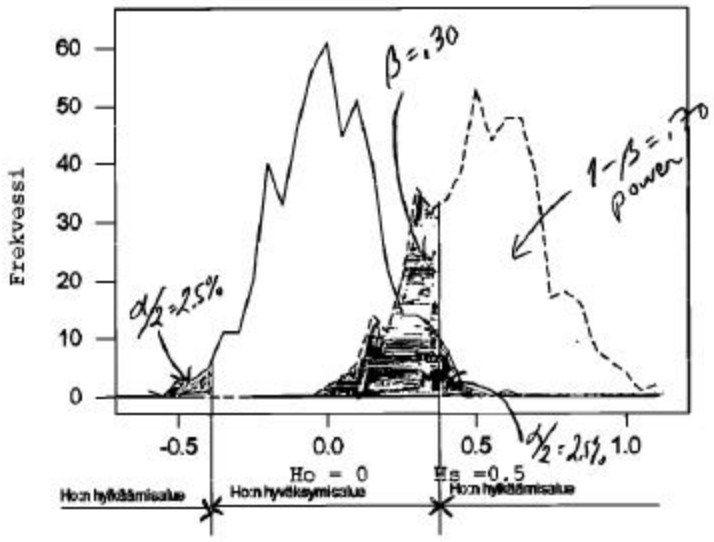
2) Erotusten otantajakauma, jos myyl=myy2 eli Ho olisi tosi.
Kuviot 2-4 perustuvat simulaatioon 500:lla erotuksella (kussakin Nl=N2=50).



3) Erotusten otantajakauma, jos Hs: myy2-myyl = 0.5 olisi tosi.



4) Kohdat 2 ja 3 yhdistettynä: Ho:n hyväksymis- ja hylkäämisalueen, alfa- ja beta-tyypin erheet.



Korrelaatiokerrointa .10 (selitysosuus 1 %) pidetään pienimpänä effektin suuruutena, jota tutkimuksessa voi järkevästi tavoitella. Korrelaatiokertoimen otantajakauman hajonta (keskivirhe) on suurilla otoksilla riittävällä tarkkuudella l/sqrt(N-3). Korrelaatiokertoimen merkitsevyyden taulukkoa voi käyttää alkeelliseen voimakkuuden arviointiin. Nollahypoteesin hyväksymis-hylkäämisrajalla (kriittinen piste kun alfa=.05) oleva Hs löytyy voimakkuudella .50.

Eli tilanteessa: N=200, Ho=perusjoukossa nollakorrelaatio, Hs=perusjoukossa korrelaatio on .14, alfa=.05, on testin voimakkuus .50. Kun Hs:ksi valitaan 1 %:n merkitsevyysrajalla oleva korrelaatio kasvaa voimakkuus noin 70 %:iin. Kun Hs arvoksi asetetaan 0.1 %:n raja-arvo eli .23 saavutetaan sen suhteen jo voimakkuus 90 %, kun muut tekijät pysyvät ennallaan.

Perinteisesti alfa-tyypin virheen täsmentämistä ja välttämistä on pidetty tärkeämpänä kuin beta-tyypin erhettä. Yhdessä ne muodostavat kuitenkin tilastollisen päätöksenteon kokonaiskuvan.

8. Yhteenvetoa merkitsevyyden testauksesta

Tilastollinen merkitsevyyden testaus on hyödyllinen apuväline, mutta samalla usein väärin käsitetty ja perustelemattomiakin johtopäätöksiä aiheuttava toimenpide. Tästä syystä on aiheellista lopuksi tehdä katsaus siihen, mitä se tekee ja mitä se ei tee sekä osoittaa tavallisimpia väärinkäsityksiä sen käytössä.

Merkitsevyyden testaus on ennen kaikkea keino objektiivisuuden lisäämiseksi tutkimuksessa. Tutkija joutuu usein tilanteeseen, jossa tuloksia tekee mieli selitellä ja tulkita sen mukaan, mitkä tulokset ovat edullisia tai haluttavia. Tutkimuksen antamat tiedot ovat harvoin niin yksiselitteisiä, etteikö niitä selittelemällä ja tulkitsemalla saataisi näyttämään enemmän tutkijan haluamilta. Tämä on tietysti epätoivottavaa ja varsinkin luonnontieteellistä tutkimusta esikuvanaan pitävä positivistinen suuntaus on hakenut keinoja, joilla tutkijasta aiheutuvaa tulkinnanvaraisuutta voitaisiin vähentää. Merkitsevyyden testaus on juuri tällainen keino. Johtopäätös lopputuloksen todellisuudesta tehdään yleisesti tunnettujen, asiantuntijoiden hallitsemien menettelytapojen mukaan, eikä tutkija voi selitellä sitä edukseen samalla tavoin kuin vapaammassa, enemmän tulkintaan perustuvassa tutkimuksessa.

Edellisestä seuraa kääntäen, että merkitsevyyden testaus ei ole keino tulosten merkittävyyden tai tärkeyden arvioimiseksi. Merkitsevä lopputulos on vain sellainen, joka on epätodennäköistä saada sattumalta. On aivan eri asia, onko tuloksella merkitystä tai käytännön sovellusarvoa. Voi olla, että erittäin merkitsevälläkään tuloksella ei ole merkitystä tai niin, että alle tilastollisen merkitsevyyden jäänyt ero tai yhteys osoittautuu erittäin merkittäväksi "heikoksi signaaliksi".

Merkitsevyyden ja merkittävyyden välinen ero johtuu osittain siitä, että merkitsevyyteen vaikuttavat muutkin seikat kuin yhteyden tai eron voimakkuus. Näistä muista seikoista on tutkittavan joukon koko, numerus, erittäin tärkeä. Aiemmin esitetystä lienee tullut selväksi, että merkitseviä tuloksia on sitä helpompi saada, mitä suurempaa joukkoa tutkitaan. Muutaman henkilön otoksessa täytyy vallita hyvin voimakas korrelaatio tai olla suuri ero ennen kuin se on merkitsevä. Tämä on helppo huomata esim. tutkimalla taulukkoa korrelaatioiden merkitsevyyk-

sistä. Kymmenen henkilön otoksessa tulee valita .63:n suuruinen korrelaatio ennen kuin se on edes 5 %:n tasolla merkitsevä. Toisaalta, .09:n korrelaatio riittää viiden prosentin tasoon, jos otos on 500 henkilön suuruinen. Tämähän merkitsee sitä, että sama ero tai yhteys voi olla tai olla olematta merkitsevä aina sen mukaan, minkä kokoinen tutkijan otos on.

Se, onko lopputulos merkitsevä, ei siis kerro sitä, onko effekti (yhteys, ero) suuri vai pieni. Joskus tutkija hakee etukäteen melko pieniksi, mutta tärkeiksi tietämiään yhteyksiä. Esimerkiksi .30 korrelaatio voi joissakin yhteyksissä olla varsin huomioonotettava. Tällöin on turha lähteä tekemään tutkimusta kahdenkymmenen henkilön otoksella; odotetun suuruinen yhteys jää joka tapauksessa alle merkitsevyystason. On siis erittäin toivottavaa muodostaa etukäteen käsitys siitä, minkä suuruisia effektejä hakee ja hankittava otos tämän mukaan. Samoin on syytä raportoida merkitsevyyden lisäksi myös effektin, siis esim. keskiarvojen eron tai korrelaation suuruus, jotta lukija voi muodostaa käsityksen tilanteesta.

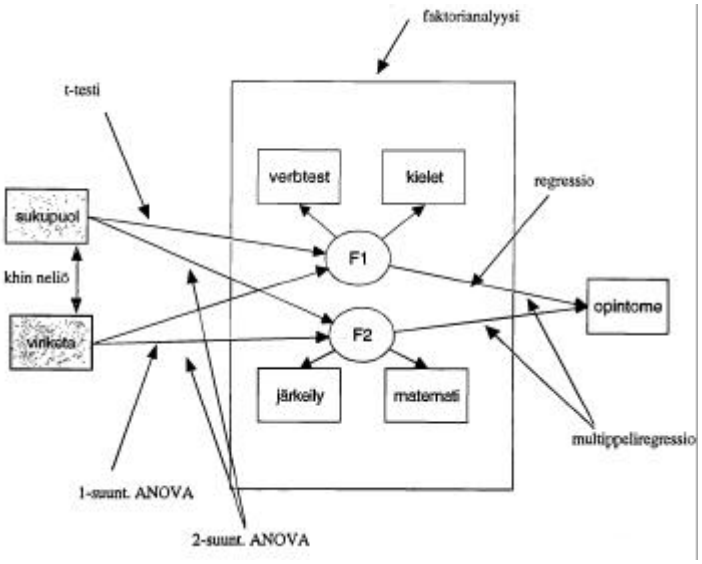
Päällisin puolin näyttää merkitsevyyden testaus erittäin täsmälliseltä toimenpiteeltä. Tämän ei pidä kuitenkaan antaa harhauttaa itseään. Toimenpiteessähän on mukana useita vaiheita, joissa suoritetaan estimointia: otantajakauman ominaisuudet arvioidaan otoksen perusteella, jakaumat oletetaan tietyn muotoiseksi, binomijakaumaa pidetään jatkuvana ja normaalina, otos oletetaan edustavaksi jne. Usein lienee parasta tyytyä näkemykseen, että merkitsevyyden testaus antaa hyvää lisäinformaatiota tutkijalle siitä, minkälaisessa tilanteessa hän toimii, mutta lopullinen päätös tehdään myös muiden seikkojen avulla. Esimerkiksi tulosten johdonmukaisuus tai selitettävyys ovat usein arvokkaita indikaattoreita niiden uskottavuuden kannalta.

Mikä kirjassa on keskeistä?

Otantajakauma ja sen käyttö tilastollisessa päättelyssä on asian ymmärtämisen kannalta eräs keskeisimpiä. Sitä on sen vuoksi esitelty hiukan harvemminkin esiintyvissä sovellustilanteissa. Keskiarvojen ja hajontatietojen käyttö kuvaamisen apuna ja ryhmävertailuissa on tärkeä. Siihen liittyy erottamattomasti yksisuuntainen varianssianalyysi (joka sisältää myös t-testin erityistapauksenaan). Yhdellä (siitä sana yksisuuntainen) luokittelevalla tekijällä tarkastellaan toista kvantitatiivista muuttujaa. Ristiintaulukoinnin käyttö ja siihen liittyvä khiin

neliö on myös tärkeä asiakokonaisuus. Useamman kuin kahden muuttujan yhteistarkastelujen perusmenetelmiä ovat regressioanalyysi (yleensä sanaa ei käytetä yhden muuttujan regressiosta toiseen vaan useamman) ja faktorianalyysi. Siinä luettelo tärkeistä asioista.

Jorma Kuusela on opetuksensa yhteydessä käyttänyt oheista tapaa vetää yhteen käytetyt menetelmät. Siinä näkyy myös osviittaa mallintamiseen. Manifestien ja latenttien piirteiden polkuanalyysiä käyttäen hahmotettu malli voitaisiin kokonaisuutena testata (SEM-ohjelmilla esim. EQS, AMOS tai LISREL). Data istuu malliin kuulemma varsin hyvin.



Tässä esitetään vain tulokset erillisistä yhteyksistä. Voit kokeilla saisitko samat tulokset omilla laskelmillasi. Aineisto on nopeasti sisään kirjoitettu tai se on saatavilla tiedostona opettajaltasi.

1) Sukupuolen ja viriketaustan khiin neliö saa arvon 0.188, p=.910 eli sukupuolen mukaan ei ole valikoiduttu erilaisista ympäristöistä. Muuttujien r=.044, p=.816 (silloin viriketaustaa pidetään kvantitatiivisena muuttujana). Sama lopputulos tässä tapauksessa. Huomaa, että r testaa keskiarvojen eroa sukupuolen

mukaan (kuten t-testi) ja khii-toiseen viriketaustan jakauman muodon eroa sukupuolen mukaan - siinä pieni, mutta joissakin yhteyksissä tärkeä ero.

2) Naisten ja miesten erot verbaalisella faktorilla ovat selvät. Eta-toiseen on 28.9 % ja p=.002. Naiset menestyvät paremmin. Päättelyfaktorilla tulos on: eta-toiseen 21.4 %ja p=.010. Asian voit tutkia korrelaationa, yksisuuntaisella varianssianalyysilla tai t -testillä. Kaikista tulee täsmälleen sama tulos (tilastollisena päättelynä).

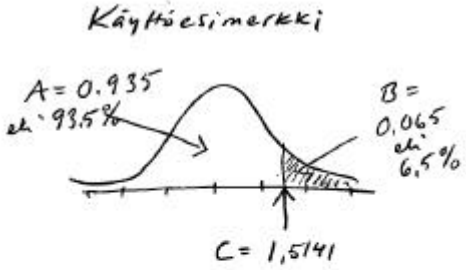
3) 2-suuntaiset (sukup ja viriket) varianssianalyysit faktoreihin tuottavat kokonaisselitykset seuraavasti. Pafl (verbaalinen) kohteena selitysaste on 80.0 %. Sekä sukupuoli (F=29.984, dfl=l, df2=24, p=.000) että viriketausta (F=29;956, df1=2, df2=24, p=.000) ovat merkitseviä selittäjiä, yhdysvaikutusta (F=0.858, dfl=2, df2=24, p=.437) ei ole. Paf2 kohteena selitysaste on 51.2 % ja p=.003. Sukupuoli (F=11.200, p=.003) ja viriketausta (F=5.551, p=.010) ovat edelliseen verrattuna heikompia mutta merkitseviä. Yhdysvaikutusta ei tilastollisesti ole (F=1.425, p=.260).

Osa edellisistä on jo kirjassa selostettukin. Faktorianalyysi ja faktoreilla tehty regressioanalyysi opintomenestykseen on selostettu oleellisin osin kirjassa.
Hyvä empiirinen tutkimus pyrkii teoriavetoiseen mallintamiseen. Kuvio antaa tuollaisen toiminnan kokonaisvaltaisuudesta ja havainnollisuudesta hyvän esimerkin. Erillisinä analyysit helposti jäävät silppumaiseksi osailmiöiden tarkasteluksi. Hyvän kvantitatiivisen tutkimuksen tekeminen saattaa olla yhtä vaikeaa kuin hyvän laadullisen tutkimuksen suorittaminen. Molemmissa on hallittava-joukko menetelmällisiä perussääntöjä (omista lähtökohdistaan). Kumpikin voi olla "sitaattikokoelmaa": toinen tietokonetulostuksesta, toinen haastattelunauhan litteroinnista. Yksityiskohtien tasolta mielenkiintoisiin, teoriavetoisiin "teemoihin" kohoaminen raportoinnissa vaati kvantitatiivisessa otteessa muutakin kuin tilastomenetelmien hallitsemisen taitoja.

Liitteet

Taulukko 1. Eri kohdista kahtia jaetun normaalijakauman suuremman osan osuus koko jakaumasta (A), pienemmän osan osuus koko jakaumasta (B) sekä jakokohtaa vastaava Z-arvo (C).

A	B	C	A	B	C	A	B	C
.500	.500	.0000	.725	.275	.5978	.950	.050	1.6449
.505	.495	.0125	.730	.270	.6128	.955	.045	1.6954
.510	.490	.0251	.735	.265	.6280	.960	.040	1.7507
.515	.485	.0376	.740	.260	.6433	.965	.035	1.8119
.520	.480	.0502	.745	.255	.6588	.970	.030	1.8808
.525	.475	.0627	.750	.250	.6745	.975	.025	1.9600
.530	.470	.0753	.755	.245	.6903	.980	.020	2.0537
.535	.465	.0878	.760	.240	.7063	.985	.015	2.1701
.540	.460	.1004	.765	.235	.7225	.990	.010	2.3263
.545	.455	.1130	.770	.230	.7388	.995	.005	2.5758
.550	.450	.1257	.775	.225	.7554	.996	.004	2.6521
.555	.445	.1383	.780	.220	.7722	.997	.003	2.7478
.560	.440	.1510	.785	.215	.7892	.998	.002	2.8782
.565	.435	.1637	.790	.210	.8064	.999	.001	3.0902
.570	.430	.1764	.795	.205	.8239	.9995	.0005	3.2905
.575	.425	.1891	.800	.200	.8416			
.580	.420	.2019	.805	.195	.8596			
.585	.415	.2147	.810	.190	.8779			
.590	.410	.2275	.815	.185	.8965			
.595	.405	.2404	.820	.180	.9154			
.600	.400	.2533	.825	.175	.9346			
.605	.395	.2663	.830	.170	.9542			
.610	.390	.2793	.835	.165	.9741			
.615	.385	.2924	.840	.160	.9945			
.620	.380	.3055	.845	.155	1.0152			
.625	.375	.3186	.850	.150	1.0364			
.630	.370	.3319	.855	.145	1.0581			
.635	.365	.3451	.860	.140	1.0803			
.640	.360	.3585	.865	.135	1.1031			
.645	.355	.3719	.870	.130	1.1264			
.650	.350	.3853	.875	.125	1.1503			
.655	.345	.3989	.880	.120	1.1750			
.660	.340	.4125	.885	.115	1.2004			
.665	.335	.4261	.890	.110	1.2265			
.670	.330	.4399	.895	.105	1.2536			
.675	.325	.4538	.900	.100	1.2816			
.680	.320	.4677	.905	.095	1.3106			
.685	.315	.4817	.910	.090	1.3408			
.690	.310	.4959	.915	.085	1.3722			
.695	.305	.5101	.920	.080	1.4051			
.700	.300	.5244	.925	.075	1.4395			
.705	.295	.5388	.930	.070	1.4757			
.710	.290	.5534	.935	.065	1.5141			
.715	.285	.5681	.940	.060	1.5548			
.720	.280	.5828	.945	.055	1.5982			



Taulukko 2. Korrelaatiokertoimen ja Fisherin Z:n muuntataulukko

r	Z_F	r	Z_F	r	Z_F	r	Z_F	r	Z_F
.00	.000	.20	.203	.40	.424	.60	.693	.80	1.099
.01	.010	.21	.213	.41	.436	.61	.709	.81	1.127
.02	.020	.22	.224	.42	.448	.62	.725	.82	1.157
.03	.030	.23	.234	.43	.460	.63	.741	.83	1.188
.04	.040	.24	.245	.44	.472	.64	.758	.84	1.221
.05	.050	.25	.255	.45	.485	.65	.775	.85	1.256
.06	.060	.26	.266	.46	.497	.66	.793	.86	1.293
.07	.070	.27	.277	.47	.510	.67	.811	.87	1.333
.08	.080	.28	.288	.48	.523	.68	.829	.88	1.376
.09	.090	.29	.299	.49	.536	.69	.848	.89	1.422
.10	.100	.30	.310	.50	.549	.70	.867	.90	1.472
.11	.110	.31	.321	.51	.536	.71	.887	.91	1.528
.12	.121	.32	.332	.52	.576	.72	.908	.92	1.589
.13	.131	.33	.343	.53	.590	.73	.929	.93	1.658
.14	.141	.34	.354	.54	.604	.74	.950	.94	1.738
.15	.151	.35	.365	.55	.618	.75	.973	.95	1.832
.16	.161	.36	.377	.56	.633	.76	.996	.96	1.946
.17	.172	.37	.388	.57	.648	.77	1.020	.97	2.092
.18	.182	.38	.400	.58	.662	.78	1.045	.98	2.298
.19	.192	.39	.412	.59	.678	.79	1.071	.99	2.647

Taulukko 3. t-arvon merkitsevyys

	Suuntaamaton testaus				
df	10%	5%	2%	1%	0.1%
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
∞	1.645	1.960	2.326	2.576	3.291
	5%	2.5%	1%	0.5%	0.05%
	Suunnattu testaus				

Taulukko 4. Korrelaatiokertoimen merkitsevyys.

df=N—2	5%	1%	0.1%
4	.811	.917	.974
6	.706	.834	.924
8	.631	.764	.872
10	.576	.707	.823
12	.532	.661	.780
14	.497	.622	.742
16	.468	.589	.708
18	.443	.561	.678
20	.422	.536	.652
25	.380	.486	.597
30	.349	.448	.554
35	.324	.418	.518
40	.304	.393	.489
45	.287	.372	.464
50	.273	.354	.443
60	.250	.324	.407
70	.231	.301	.379
80	.217	.283	.356
90	.205	.267	.337
100	.194	.254	.321
200	.139	.182	.233
500	.088	.115	.150

Taulukko 5. Khiin neliön merkitsevyys.

df	5%	1%	0.1%
1	3.8	6.6	10.8
2	6.0	9.2	13.8
3	7.8	11.3	16.3
4	9.5	13.3	18.5
5	11.1	15.1	20.5
6	12.6	16.8	22.5
7	14.1	18.5	24.3
8	15.5	20.1	26.1
9	16.9	21.7	27.9
10	18.3	23.2	29.6
11	19.7	24.7	31.3
12	21.0	26.2	32.9
13	22.4	27.7	34.5
14	23.7	29.1	36.1
15	25.0	30.6	37.7
16	26.3	32.0	39.3
17	27.6	33.4	40.8
18	28.9	34.8	42.3
19	30.1	36.2	43.8
20	31.4	37.6	45.3
21	32.7	38.9	46.8
22	33.9	40.3	48.3
23	35.2	41.6	49.7
24	36.4	43.0	51.2
25	37.7	44.3	52.6
26	38.9	45.6	54.0
27	40.1	47.0	55.5
28	41.3	48.3	56.9
29	42.6	49.6	58.3
30	43.8	50.9	59.7

Taulukko 6. F-suhteen merkitsevyysrajat

df1=osoittajan vapausasteet (ryhmien lukumäärä - 1)
df2=nimittäjän vapausasteet (N-g)
% = merkitsevyystaso

F-taulukko alkaa

		df_1									
df_2	%	1	2	3	4	5	6	8	12	24	∞
1	0.1	405284	500000	540379	562500	576405	585937	598144	610667	623497	636619
	1.0	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
	5.0	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	254.32
2	0.1	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.5	999.5
	1.0	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
	5.0	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	0.1	167.5	148.5	141.1	137.1	134.6	132.8	130.6	128.3	125.9	123.5
	1.0	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
	5.0	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	0.1	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41	45.77	44.05
	1.0	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
	5.0	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	0.1	47.04	36.61	33.20	31.09	29.75	28.84	27.64	26.42	25.14	23.78
	1.0	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
	5.0	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	0.1	35.51	27.00	23.70	21.90	20.81	20.03	19.03	17.99	16.89	15.75
	1.0	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
	5.0	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	0.1	29.22	21.69	18.77	17.19	16.21	15.52	14.63	13.71	12.73	11.09
	1.0	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
	5.0	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	0.1	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.19	10.30	9.34
	1.0	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
	5.0	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93

jatkuu

df_2	%	1	2	3	4	5	6	8	12	24	∞
9	0.1	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57	8.72	7.81
	1.0	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
	5.0	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	0.1	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.45	7.64	6.76
	1.0	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
	5.0	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	0.1	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63	6.85	6.00
	1.0	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
	5.0	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	0.1	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.00	6.25	5.42
	1.0	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
	5.0	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	0.1	17.81	12.31	10.21	9.07	8.35	7.86	7.21	6.52	5.78	4.97
	1.0	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
	5.0	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	0.1	17.14	11.78	9.73	8.62	7.92	7.43	6.80	6.13	5.41	4.60
	1.0	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
	5.0	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
16	0.1	16.12	10.97	9.00	7.94	7.27	6.81	6.19	5.55	4.85	4.06
	1.0	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
	5.0	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
18	0.1	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13	4.45	3.67
	1.0	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
	5.0	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
20	0.1	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
	1.0	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
	5.0	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84

jatkuu

df_2	%	1	2	3	4	5	6	8	12	24	∞
25	0.1	13.88	9.22	7.45	6.49	5.88	5.46	4.91	4.31	3.66	2.89
	1.0	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
	5.0	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
30	0.1	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
	1.0	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
	5.0	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	0.1	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.64	3.01	2.23
	1.0	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
	5.0	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	0.1	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.31	2.69	1.90
	1.0	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
	5.0	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
∞	0.1	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00
	1.0	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00
	5.0	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

F-taulukko päättyy